RMP Colloquia

This section, offered as an experiment beginning in January 1992, contains short articles intended to describe recent research of interest to a broad audience of physicsts. It will concentrate on research at the frontiers of physics, especially on concepts able to link many different subfields of physics. Responsibility for its contents and readability rests with the Advisory Committee on Colloquia, U. Fano, chair, Robert Cahn, S. Freedman, P. Parker, C. J. Pethick, and D. L. Stein. Prospective authors are encouraged to communicate with Professor Fano or one of the members of this committee.

The single-electron transistor

M. A. Kastner

Department of Physics and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

The discovery of periodic conductance oscillations as a function of charge density in very small transistors has led to a new understanding of the behavior of electrons in such small structures. It has been demonstrated that, whereas a conventional transistor turns on only once as electrons are added to it, submicron-size transistors, isolated from their leads by tunnel junctions, turn on and off again every time an electron is added. This unusual behavior is primarily the result of the quantization of charge and the Coulomb interaction between electrons on the small transistor. However, recent experiments demonstrate that the quantization of energy is important as well.

CONTENTS

I. INTRODUCTION

In 1978 a new subfield of condensed matter physics was born when Thouless (1977) pointed out that the size of a conductor, if made small enough, plays a role in determining its electronic properties. In the following decade a variety of theoretical and experimental discoveries were made in what has come to be called mesoscopic physics. The name was chosen to describe submicron-size systems in which the behavior of electrons can be fundamentally different from that in atoms, which are much smaller, and from that in macroscopic samples, which are much larger. An introduction to the ideas of mesoscopic physics may be found in the article in *Physics Today* by my colleagues, Altshuler and Lee (1988). In this Colloquium I shall describe a series of experiments we have done on mesoscopic transistors. I shall show that when a transistor is small enough, say a few thousand angstroms or less in all dimensions, it behaves in a very unusual way. Whereas a conventional transistor turns on only once when electrons are added to it, a very small transistor turns on and off again every time an electron is added to it.

This single-electron transistor was an accidental discovery we made several years ago. Since then we have found a way of controlling its behavior, and in the past year a quite simple model for the underlying physics has emerged. In Sec. II, I tell the story of our initial discovery that the conductance of small transistors turns on and off periodically as a function of the electron density. In Sec. III, I show how we demonstrated that the period corresponds to the addition of a single electron to the transistor. In Sec. IV, I present a simple model, in which the oscillatory conductance results from the Coulomb interaction between electrons on the transistor and the quantization of charge, but in which the mechanics are entirely classical. In order to understand the details of the oscillatory behavior, however, and especially the phenomena we observe in high magnetic fields, we need the quantization of energy, as well as that of charge. I discuss these effects in Sec. V. I want to emphasize at the outset, as I shall again in the conclusion in Sec. VI, that our journey of exploration is not yet complete and that these transistors continue to surprise us.

II. THE DISCOVERY OF CONDUCTANCE OSCILLATIONS IN SMALL TRANSISTORS

Since the early 1980s our group[1] had been studying the properties of very narrow transistors at low tempera-

[1]In addition to myself, our group includes M. Heiblum and S. Wind at IBM, H. I. Smith, and D. A. Antoniadis at MIT, whose technical expertise and sophisticated ideas were essential to success. Even more important are the young people who really made it happen: The initial discovery was made by a graduate student, John Scott-Thomas, and postdoctoral associate, Stuart Field. The evidence for single-electron oscillations and all the recent results came from devices invented by another student, Udi Meirav, and the high-magnetic-field experiments were done by Paul McEuen, a postdoctoral associate, and Ethan Foxman, a graduate student. We have also benefitted greatly from stimulating interaction with two postdoctoral associates in the MIT condensed-matter theory group, Ned Wingreen and Yigal Meir.

tures, about 100 mK (Kastner *et al.*, 1987, and references therein). In general, such devices show reproducible noiselike variations of their conductance as a function of their electron density. When the transistors are sufficiently conducting, the magnitudes of these fluctuations are always of order e^2/h, as is the case for fluctuations observed as a function of magnetic field in narrow metal wires. Because the magnitude is always the same, they are called universal conductance fluctuations. In 1988 John Scott-Thomas succeeded in fabricating a new kind of narrow transistor (Scott-Thomas *et al.*, 1989; Field *et al.*, 1990). We expected to see similar fluctuations in this kind of device, but, instead, we saw something very different. Before describing the surprising discovery we made, I need to tell you how the new transistor works.

The structure Scott-Thomas made is similar in many ways to the metal-oxide-semiconductor field-effect transistor (given the acronym MOSFET), which is the device commonly used in computer memories. A MOSFET is a parallel-plate capacitor, one plate of which is a metal and the other of which is a semiconductor, in this case silicon. When a positive voltage is applied to the metal, electrons accumulate in the conduction band of the semiconductor. The electric field at the interface between the Si and the SiO_2, which is the insulator of the capacitor, is so strong that at low temperatures the electrons are confined to move in only two dimensions, parallel to the interface. The conductance of this two-dimensional electron gas, measured by means of two (n-type) contacts, increases as the positive voltage on the gate is raised because the number of electrons on the semiconductor increases. It was in such a two-dimensional electron gas that the quantum Hall effect was discovered.

Our device, for which a schematic diagram is given in Fig. 1(a), is different only in that it has two metal gates instead of one: The top gate is continuous as in the conventional MOSFET, but the lower gate has a narrow gap in it, about 700 Å wide. The small size of this gap is what makes fabrication challenging. X rays or electrons must be used to define this narrow channel because light, which is currently used to make integrated circuits, has wavelengths too large to make structures so small. I do not have the space to describe adequately the fascinating technology invented to make these nanometer-size structures, but you can find descriptions in the references (Scott-Thomas *et al.*, 1989; Field *et al.*, 1990).

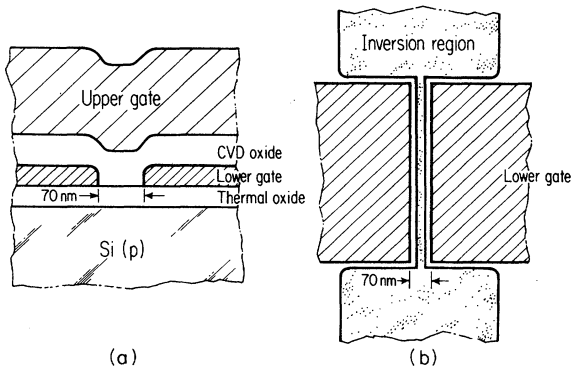

FIG. 1. Schematic (a) cross section and (b) top view of the silicon transistor with continuous upper gate and a gap in the lower gate. The electron gas, formed in the Si by the positively biased upper gate, is confined by the lower gate. The Si is p-type, so the surface electrons are isolated from the bulk by p-n junctions. That is why the electron-rich region is called an inversion layer. The cross section is roughly to scale, but the top view is not. The narrow channel is typically 20 nm wide by 1–10 μm long. Contact is made to the two wide inversion regions.

When the upper gate is raised to a voltage that is positive with respect to the semiconductor, and the bottom gate is neutral or negative, electrons are added to the semiconductor only under the gap in the bottom gate. This creates an electron gas that is confined to move in only one direction [see Fig. 1(b)]. We had made one-dimensional transistors before. The only difference between this narrow transistor and those we had made earlier was that it was narrower, by a factor of 3 to 5 (the actual width of the electron gas was 15 to 30 nm, narrower than the gap in the lower gate), and that the electrons in it had higher mobility, an indication that there were fewer charges at the interface to scatter electrons.

We expected this transistor to behave like earlier ones, which displayed the random universal conductance fluctuations. However, as shown in Fig. 2, the behavior was completely different. What is plotted in the upper panel is the conductance of the narrow gas of electrons as a function of the gate voltage V_g. Remember that the number of electrons per unit length added is just the capacitance per unit length times the change in V_g, i.e., $e\Delta(N/L)=(C/L)\Delta V_g$, so data like those in Fig. 2 can be

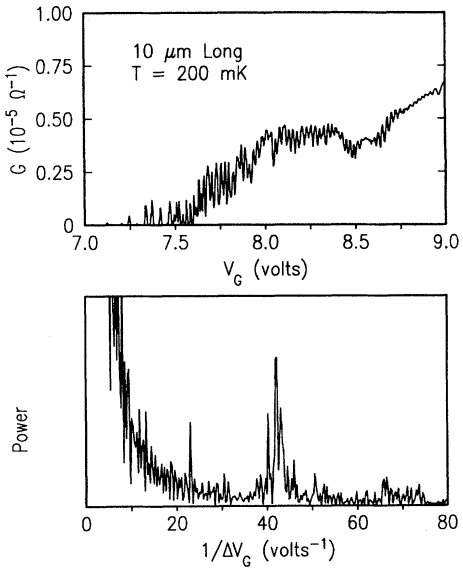

FIG. 2. One of the first measurements on a narrow Si transistor in which periodic oscillations were observed. Although the oscillatory behavior is not obvious from the conductance G itself (upper curve) as a function of V_g, the Fourier transform shows a large peak.

thought of as conductance versus electron density, N/L.

At first glance, the structure in the data of Fig. 2 looked very similar to the random fluctuations we had seen before. On closer inspection, however, it became clear that these fluctuations were *periodic*. This was quite a surprise. To convince ourselves that what we were seeing were really periodic oscillations and not universal conductance fluctuations, we calculated the Fourier transform of the data, and that is shown in the lower part of Fig. 2. After looking at many samples, we found one that behaved as shown in Fig. 3, for which it is really unnecessary to resort to the Fourier transform, although that is shown in the lower part of Fig. 3 for completeness.

Having observed that our transistor's conductance oscillated as a function of the number of electrons per unit length, we saw that there was a special length in the problem: The voltage difference for one period is that necessary to add some fixed number of electrons to some fixed length, that is, $e\Delta N = L_0(C/L)\Delta V_g$. It was natural to guess that the length L_0 was that of the transistor itself. However, as can be seen in Fig. 4, that guess was wrong. The oscillations are shown there for three different lengths; the period was not correlated with the sample length at all.

In fact, the period changed every time we cycled a specific sample to room temperature and back to low temperature again, as shown in Fig. 5, where data are given for a single sample at two times, six months apart. Every time we cooled the transistor we saw oscillations, and the period was the same order of magnitude, but the exact value of the period was random. I should emphasize that the data, like those in Figs. 2–5, were reproducible, in detail, if we kept the device cold.

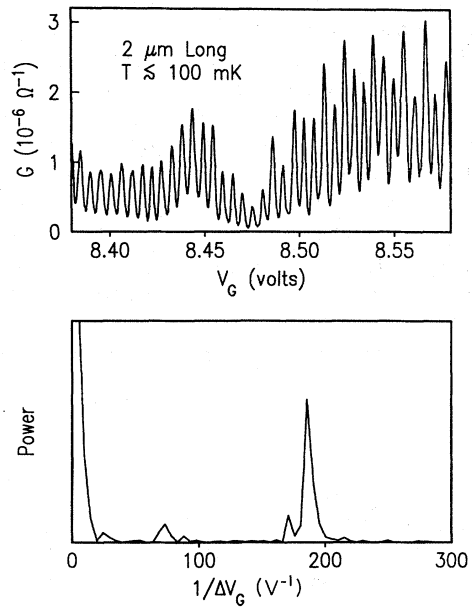

FIG. 3. Conductance G vs V_g for the most obvious oscillatory behavior seen in Si devices. The Fourier transform is also shown.

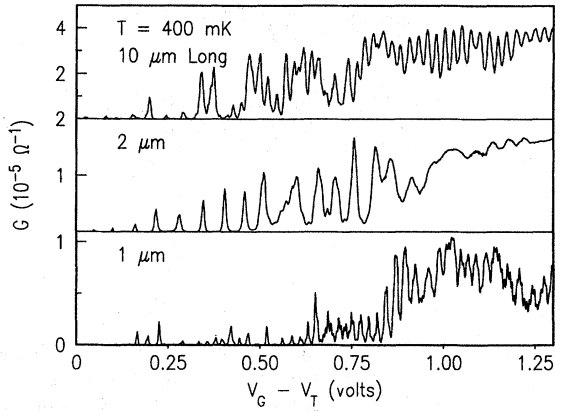

FIG. 4. G vs V_g for Si devices with 10-, 2-, and 1-μm-long channels, from top to bottom. Clearly, the period is uncorrelated with the channel length.

This sensitivity to thermal cycling was something we had seen before in studying universal conductance fluctuations, so we recognized that the period was determined by the distribution of charges at the Si-SiO_2 interface. It is well known that such charges are always present, at densities of about 10^{10} per cm^2 for the best interfaces, so a transistor, like ours, which is $\sim$20 nm$\times$1 μm in size, has, on average, about two of these charges adjacent to it. Every sample has a different random distribution of charges and, furthermore, every time the sample is warmed to room temperature the distribution of charges changes. We guessed that the charges create potential barriers along the length of the transistor, as illustrated schematically in Fig. 6. We now had a postulate that explained the data: Each period of the conductance oscillations corresponded to the addition of one electron to the distance between the charges.

This was just a guess, however. How could we know that the period does not correspond to the addition of

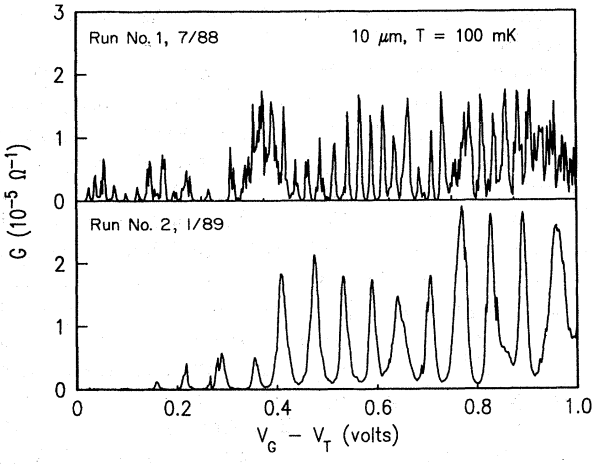

FIG. 5. Two measurements of G vs V_g for the same 10 μm Si device, cycling the device to room temperature and back to 100 mK between measurements. The data in the top panel were taken six months before those in the lower panel.

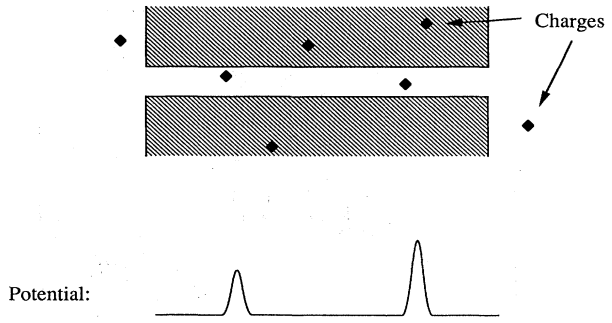

FIG. 6. Sketch (above) of narrow channel through which electrons move. The channel is determined by a metal gate with gap, as indicated by the shaded area. The black diamonds indicate typical random distribution of charges near the interface. Sketch (below) of electrostatic potential resulting from such impurities vs position along the channel.

two electrons because of the spin degeneracy? In fact, Si has an additional degeneracy because of its energy-band structure, so the period might require *four* electrons. We would never know the answer unless we could measure the length L_0 directly.

III. SINGLE-ELECTRON CHARGING

At the time of our discovery of oscillations in Si transistors, Udi Meirav, working with Shalom Wind and Moti Heiblum at IBM, had just succeeded in making closely analogous small transistors in GaAs (Meirav *et al.*, 1989; for technical details see also Meirav *et al.*, 1988, and Wind *et al.*, 1990). One of the advantages of this material is that the density of charges near the interface between the semiconductor and insulator is smaller than for the Si-SiO_2 case. Meirav realized that this made it possible to induce potential barriers intentionally (Meirav, Kastner, and Wind, 1990), so the distance between them would be known. The way he did this is illustrated in Fig. 7.

One begins with a heavily doped crystal of GaAs (labeled n^+). This plays the role of the uniform *top* metal gate in the Si transistor. Next, using molecular-beam epitaxy (MBE), one grows a layer of AlGaAs, which is a semiconductor with a larger band gap than GaAs. This is the insulator, like the SiO_2 in the Si structure (Fig. 1). Next, one grows a layer of pure GaAs, which is where the electrons accumulate, as in the silicon for the earlier structure. A positive voltage applied to the n^+ substrate controls the density of these electrons. Finally, a metal gate is deposited on the top and patterned using electron-beam lithography. This metal is negatively biased, so electrons are repelled from it, confining the electrons to a narrow channel, just as the gate with a gap (Fig. 1) confines the electrons in the Si transistor.

Making a GaAs transistor work just like the Si devices was already a major technological achievement. However, the special new feature that opened unexpected ave-

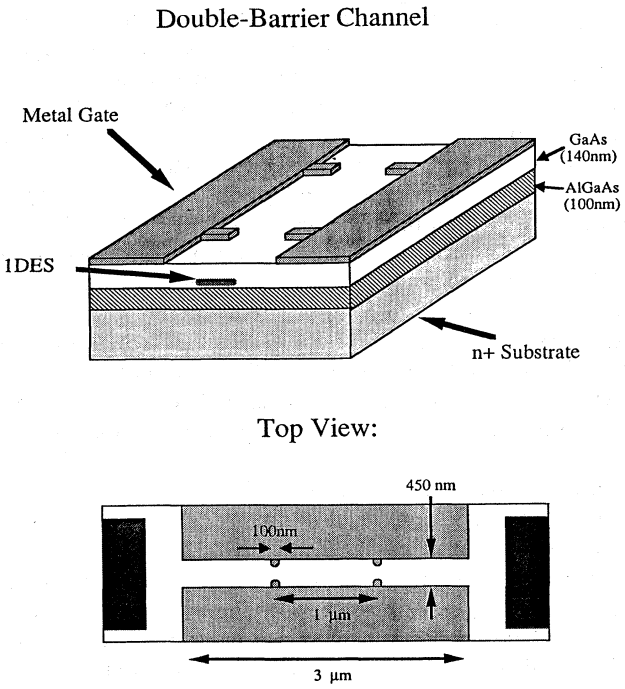

FIG. 7. Schematic drawings of device structure. Top: A one-dimensional electron gas (1DES) or narrow two-dimensional gas forms at the *top* GaAs-AlGaAs interface, with a density controlled by the substrate voltage V_g. Bottom: A top view showing typical dimensions of the top metal gate structure, which defines a narrow channel with two constrictions.

nues of research was the patterning of the upper metal gates to form constrictions in the channel through which the electrons flow. The idea was that the negative bias on the top gates would create a potential barrier for electrons moving down the narrow channel.

Figure 8 shows how two of the new devices, which had the same distance between the constrictions ($L_0 = 1$ μm), behaved. As before, the conductance oscillated, but with an important difference: while the amplitudes of conductance peaks were different for the two devices, the *period* of the oscillations was the *same*. Cycling to room temperature and back changed the amplitudes of conductance peaks, but left the period of oscillation unaffected. For the first time the period of the oscillations was repro-

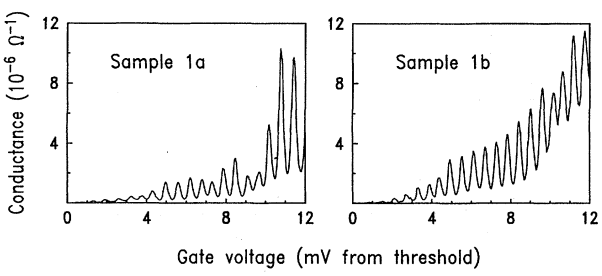

FIG. 8. Conductance as a function of V_g for two samples with the same geometry. Although the amplitudes of the conductance peaks vary, the period of the oscillations is the same.

ducible in the same transistor and for two identical transistors.

As can be seen in Fig. 9, the periodicity depends on the distance between the constrictions in the way we would expect. Shown there are data for two samples (2 and 3), which have the same width as the two in Fig. 8 but with $L_0=0.8$ μm and 0.6 μm, respectively. As expected, shorter segments have longer periods in V_g. Clearly this is consistent with the idea that each period corresponds to the addition of the same number of electrons to the region between the constrictions: For a shorter segment the capacitance is smaller, so the voltage to add one electron is larger. However, to demonstrate that the number of electrons added per period is one and not two, for example, we need to know the absolute capacitance of the segment.

To obtain estimates of the potential distribution and of the capacitance between the substrate (the gate) and the region of the electron gas delimited by the constrictions, we needed to solve the Poisson equation and find the charge density self-consistently. Arvind Kumar, an MIT student working under the guidance of Frank Stern at IBM, did this numerically (Meirav *et al.*, 1991; Laux *et al.*, 1988; Kumar *et al.*, 1989). The technique he used was self-consistent between the charge and the potential, but the charge density was estimated in a way analogous to the Thomas-Fermi approximation.

In Fig. 10 the potential is plotted as a function of position along the channel, which has constrictions at ± 0.5 μm. At $V_g=180$ mV the minimum in the potential between the constrictions drops just below the Fermi energy, which is at zero in the figure, and electrons begin to accumulate in the isolated region. As the gate voltage is increased further, the barriers drop rapidly, so that above $V_g=300$ mV the Fermi energy is higher than the barriers. The conductance oscillations occur in the limited range of voltage in which there is an electron gas isolated from the leads by tunnel barriers.

The capacitance was obtained at each gate voltage by integrating the charge contained in the region between the two constrictions. Although the semiclassical approximation and a full quantum-mechanical treatment differ qualitatively in the potential and charge distributions, experience shows (Laux *et al.*, 1988; Kumar *et al.*, 1989) that the integrated charge, and hence the capacitance, is, to a good approximation, nearly the same in both approaches. By comparing the calculated capacitance to the measured period we find that the period of oscillations corresponds to the addition of *one* electron per oscillation. This is a truly amazing result (Meirav *et al.*, 1990): The transistors, for which data are shown in Fig. 9, turn on and off again every time an electron is added.

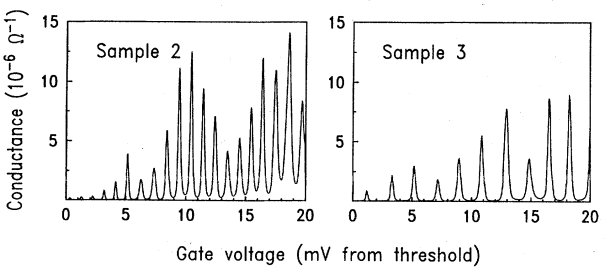

FIG. 9. Conductance as a function of V_g for two devices with different lengths: Sample 2 has $L_0=0.8$ μm, and sample 3 has $L_0=0.6$ μm. Note that samples 1(a) and 1(b) in Fig. 8 have $L_0=1.0$ μm. The period increases inversely with L_0.

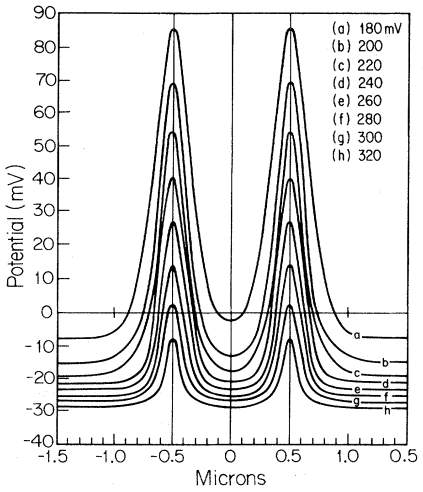

FIG. 10. Calculated potential as a function of position along the center of the conducting channel. The constrictions are at ± 0.5 μm, consistent with the $L_0=1$ μm length of samples 1(a) and 1(b) in Fig. 8. The potential is shown for several values of V_g between (a) 180 mV and (h) 320 mV. The Fermi energy is taken to be zero.

IV. THE COULOMB ISLAND

There is a very simple model that explains why a small region of electron gas, separated by tunnel junctions from its leads, has a conductance that oscillates with density. It is sometimes called the Coulomb blockade model, and was developed to explain the behavior, seen in the 1960s (Zeller and Giaever, 1969), of tunnel junctions with small metal particles inside the tunnel barrier. The theory has become quite sophisticated,[2] and I give here only a highly simplified version.

For a tunnel barrier containing a metal particle, as sketched in Fig. 11, one might have expected that electrons could tunnel from one plate of the capacitor onto the metal particle and subsequently to the other plate of the capacitor. However, because of the Coulomb interaction between electrons residing on the particle this does

[2]Kulik and Shekhter, 1975; Likharev, 1988; Amman *et al.*, 1989. For a recent review see also Averin and Likharev, 1991. Important new developments can be found in Glazman and Shekhter, 1989, and Averin and Nazarov, 1990.

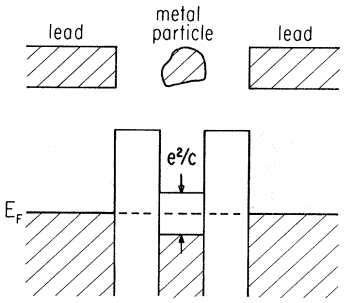

FIG. 11. Sketch of Coulomb blockade system. Top: Electrons tunnel from one lead onto a small metal particle and subsequently off the particle onto the other lead. Bottom: Energy-level spectrum for this system, showing filled (hatched) and empty (shaded) levels. The metal particle has a gap of width e^2/C in its tunneling density of states.

not happen. Current flow requires the addition of an electron or hole to the particle. To add a charge Q to the particle costs the energy $Q^2/2C$, where C is the total capacitance between the particle and the rest of the universe. Adding an electron therefore requires a Coulomb energy $e^2/2C$. There is, therefore, an energy gap in the tunneling density of states (what condensed-matter physicists call the single-particle density of states), as shown schematically in Fig. 11. For an electron to tunnel onto the particle it must have an energy above the Fermi energy in the contact by $e^2/2C$, and for a hole to tunnel it must have an energy below the Fermi energy by the same amount, so the gap has width e^2/C.

Our structure is more interesting than the metal particle in the tunnel junction because with the gate voltage we can vary the average number of electrons in the region isolated by the constrictions. If the potential difference between the gate and electron gas is V_g, the electrostatic energy of charge Q on the isolated region of the transistor is given by

$$E=-QV_g+Q^2/2C\ . \tag{1}$$

The first term is the attractive interaction between the positively charged gate electrode and the charge on the isolated region, and the second term is the repulsive interaction among the elements of charge on the isolated region. Equation (1) can be rewritten as $E=(Q-Q_0)^2/2C$, to within an additive constant, where $Q_0=CV_g$.

We can choose any value of Q_0, the charge that minimizes the energy, by varying V_g. However, because the charge is quantized, only discrete values of the energy will be possible for a given Q_0. This is illustrated in Fig. 12 for two cases. When $Q_0=Ne$, for which an integer number of electrons minimizes E, the Coulomb interaction results in the energy difference $e^2/2C$ for increasing or decreasing N by one. Under these circumstances there will be an activation energy $e^2/2C$ for current to flow as shown in Fig. 11. However, when $Q_0=(N+\frac{1}{2})e$, the state with $Q=Ne$ and that with $Q=(N+1)e$ are degenerate; the charge fluctuates between the two values even at zero temperature, so the energy gap in the tunneling density of states disappears.

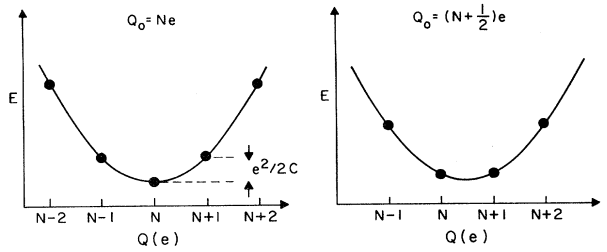

FIG. 12. Energy vs charge on isolated particle for an integer average number of electrons (left) and for a half-integer average number (right). Only those values of energy denoted by solid circles are allowed by the quantization of charge.

This simple classical model was first used by van Houten and Beenakker (1989) to explain the most obvious features of the behavior of our devices: The conductance is thermally activated at all values of the gate voltage except those for which the average charge on the isolated segment is $(N+\frac{1}{2})e$. This results at low temperature in a conductance consisting of periodic sharp peaks. The period in V_g is predicted to be simply e/C, the voltage change necessary to alter Q_0 from $(N+\frac{1}{2})e$ to $(N+\frac{3}{2})e$. The activation energy at the minimum is, indeed, found to be of order e multiplied by the period in V_g, that is, $e^2/2C$.

As we shall see shortly, this is not the whole story. However, it is important to emphasize that the Coulomb interaction between the electrons on the isolated segment is the largest energy in the problem. Because the time for tunneling onto and off of the segment is long, the number of electrons on the segment is quantized. This quantization of charge, together with the Coulomb interaction, suppresses charge fluctuations completely at zero temperature for all values of Q_0 not equal to $(N+\frac{1}{2})e$, and for the latter the charge fluctuates only by one electron. The suppression of the charge fluctuations by the Coulomb interaction is what makes the behavior of our structure so unusual, and that is why we call it a Coulomb island. The charge fluctuations at finite temperature (see references in footnote 2) and are interesting, but are not important for the present discussion.

At this point I have explained the essence of how the single-electron transistor works using a classical argument, which requires only the quantization of charge. The rest of my colloquium shows where quantum mechanics becomes important.

V. LEVEL SPECTROSCOPY OF A COULOMB ISLAND

In Figs. 8 and 9 it can be seen that the size of the conductance peaks varies, apparently at random. As recognized by Meir, Wingreen, and Lee (1991) and in parallel by Beenakker and co-workers (Beenakker, 1991;

Beenakker *et al.*, 1991), this is one manifestation of the quantization of the energy of eigenstates in the isolated region, in addition to the quantization of charge. Because the area of the device between the two constrictions is small, the energy separation of the eigenstates is appreciable, and this affects the tunneling. Figure 13 illustrates the spectrum of the single-particle density of states for a Coulomb island with a capacitance C. For the case $Q_0 = Ne$ there is a gap of width e^2/C, as for the classical case. However, the small size of the structure makes the possible values of the energy outside this gap discrete.

For any conductance peak the value of Q_0 is equal to $(N+\frac{1}{2})e$, and this requires that the Fermi energy in the leads be equal to that of one of the discrete energy levels of the small structure, as illustrated in Fig. 13. Only if these energies are the same can the charge on the structure fluctuate at zero temperature, and this is the condition that determines the gate voltage at which a peak occurs. At low temperature the conductance of a specific peak depends on the tunneling matrix element for one particular energy level of the small structure. Because this matrix element depends exponentially on the decay length of the wave function through the barriers, the amplitudes of the peaks vary appreciably from one to the next. Because of potential fluctuations caused by defects or impurities, these amplitudes fluctuate randomly while generally increasing with gate voltage as the Fermi energy approaches the top of the barriers (see Fig. 10).

Not only the amplitude, but also the position, of the conductance peaks depends on the discrete levels of the structure. The matching of the Fermi energy with an individual level at a conductance peak requires that the gate voltage difference between two adjacent peaks be e/C, as before, *plus* the energy spacing of the levels of the structure, ΔE. That is,

$$e\Delta V_N = e^2/C + \Delta E_N \ , \tag{2}$$

where ΔV_N is the voltage spacing between the $(N-1)$th and Nth peaks. This, in turn, means that the peak positions in gate voltage are directly proportional to the energy-level splittings and can be used to measure the level spectrum.

The level spectrum is, in general, quite complicated because the geometry of the isolated segment of the transistor is not highly symmetric. However, in a high magnetic field the spectrum becomes much simpler. Therefore, as Paul McEuen immediately realized, the best test of the theory would come from the behavior of our single-electron transistor in a high magnetic field.

In a two-dimensional electron gas at a high magnetic field B, the energy spectrum consists of Landau levels. The energies are given by $(n+1/2)\hbar\omega_c$, where the cyclotron frequency is $\omega_c = eB/m^*c$, and m^* is the effective mass. These energies are increased by the electrostatic confinement near the edges of the two-dimensional gas. Figure 14 shows schematically, as solid curves, how the energies of the states in the three lowest Landau levels depend on position. Near the edge of the sample all the occupied Landau levels cross the Fermi energy. The states at the Fermi energy are the edge states, sometimes described as classical skipping orbits, that play a central role in the quantum Hall effect. The states within each Landau level have discrete energies because of the finite size of the system, and when the size of the confined region is small the energy separation between quantized edge states is large. These discrete states are shown schematically by the points on the three curves in Fig. 14.

Each of the discrete states illustrated in Fig. 14 occupies an area given roughly by Φ_0/B, where Φ_0 is the quantum of flux, hc/e. As B increases, this area shrinks, the positions of the discrete states move from the edges toward the center in Fig. 14, and the degeneracy of the Landau levels thus increases with magnetic field. This causes electrons gradually to fall from higher into lower Landau levels, in the following way: Consider the case in which the magnetic field is chosen such that only the lowest two Landau levels are occupied and the Fermi energy is near the bottom of the second Landau level, as in-

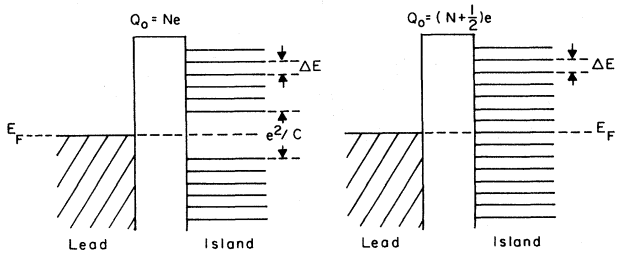

FIG. 13. Schematic energy diagram for tunneling into a Coulomb island, when the average charge on the island is Ne (left) and when it is $(N+\frac{1}{2})e$ (right). The energy levels for electronic excitation internal to the island are spaced by a small energy ΔE. The energy to add an extra electron from the leads is $e^2/2C$ when $Q_0 = Ne$. For our devices, $\Delta E \sim 0.1e^2/2C$.

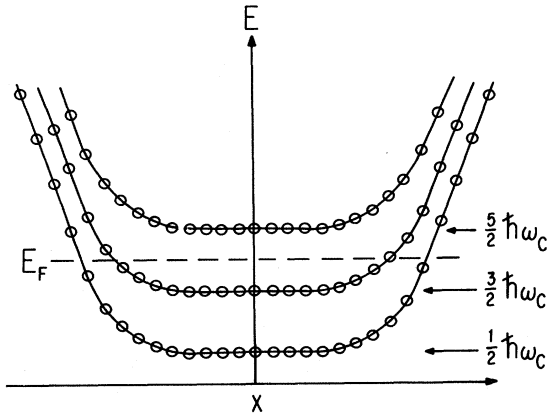

FIG. 14. Schematic of energy as a function of position in a confined geometry in a high magnetic field. The solid curves are the Landau levels, whose energies increase near the edge because of the electrostatic confinement. The circles indicate states allowed when the area of confinement is small.

dicated in Fig. 14. The quantized states near the Fermi energy, the quantized edge states, for the second Landau level move *up* in energy with increasing B, because when the Fermi energy is close to the bottom of the Landau level the increase in $\hbar\omega_c$ is more important than the increase of degeneracy. However, the quantized states near the Fermi energy in the lowest Landau level move *down* because the positions of the discrete states move toward the center, and, since the dependence of energy on position is strong at the edges, the resulting decrease in energy is more rapid than the increase of $\hbar\omega_c$. Because filled states in the second Landau level move up and empty states in the first move down, electrons fall, one at a time, from edge states in the upper Landau level into edge states in the lower Landau level as B is increased.

To make this more precise, one must choose a specific model for the electrostatic confinement, although, in the end, the results depend only weakly on the geometry. For the case of harmonic confinement the calculation was done analytically a long time ago (Fock, 1928; Darwin, 1930). For more realistic models numerical calculations are available (Sivan and Imry, 1988; Kumar *et al.*, 1990). Figure 15 shows, schematically, the trajectory of the edge states near the Fermi level for the two Landau levels in our structure. One may think of these as the contours of maximum probability for one of the quantized edge states in each Landau level. Also shown is part of the calculated energy spectrum as a function of

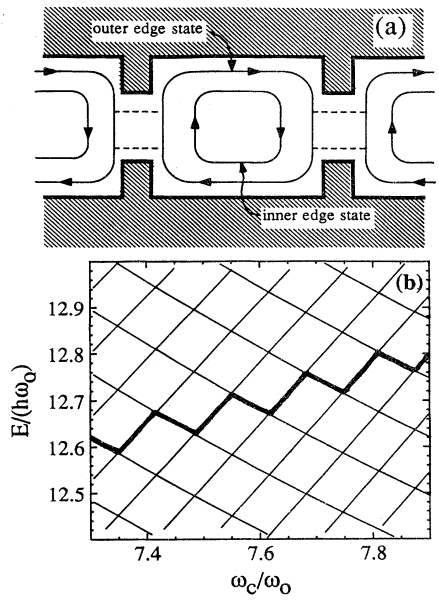

FIG. 15. (a) Schematic top view of the device, showing the path of the edge states associated with the lowest two Landau levels. The upper gate (shaded) defines a dot whose lithographic dimensions are 500 nm by 700 nm. (b) Energy level of a dot with a parabolic confining potential $\frac{1}{2}m^*\omega_0^2 r_0^2$ as a function of $\omega_c = eB/m^*c$ in a parameter range where two Landau levels are occupied (Fock, 1928; Darwin, 1930). The heavy line represents the energy of the single-particle state that is 78th lowest in energy.

B for harmonic confinement with frequency ω_0. The energies of the quantized edge states for the first Landau level increase with B and those for the second decrease with B, as expected from the qualitative discussion given above.

Figure 15 allows one to predict how the gate voltage of a conductance peak should depend on B. For a given peak in conductance the average number of electrons in the isolated segment is fixed at $N+\frac{1}{2}$. This means that the Fermi energy in the leads is degenerate with one of the quantized edge states, say one in the first Landau level, whose energy decreases with B. As a result, the gate voltage at which the peak occurs will decrease with B. This state has an average occupancy of $\frac{1}{2}$; higher energy states are empty and lower ones are full. At some field, however, the energy of a quantized edge state from the second Landau level, one that is increasing with B, crosses that of the edge state of the first level. Then the downward-moving state becomes full, the upward moving one becomes half-full, and the conductance peak, now following the state from the second Landau level, shifts to higher gate voltage. In Fig. 16, we show a measurement of the evolution of the position of one conductance peak with B. Indeed, the peak position moves alternately up and down, as expected.

Because the edge states from the first Landau level are closer to the periphery of the isolated region, the rate of tunneling into them from the leads is much higher than into those from the second Landau level. This explains the dramatic decrease in the peak height, seen in Fig. 16. Whenever the peak is shifting to higher gate voltage it is tracking a state in the second Landau level, which is more weakly coupled to the leads.

Following a single conductance peak thus gives information about the Nth state of the isolated region. By following successive peaks we can map out the entire spectrum. Remember [see Eq. (2)] that the spacing in gate

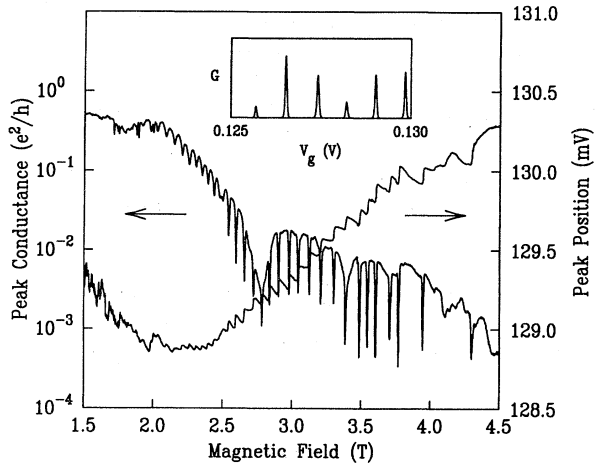

FIG. 16. Height and position of one of the conductance peaks shown in the inset as a function of magnetic field. The temperature of the electron gas is ~100 mK. The inset shows conductance vs V_g for $B=3$ T. Full scale for the inset is $0.03e^2/h$.

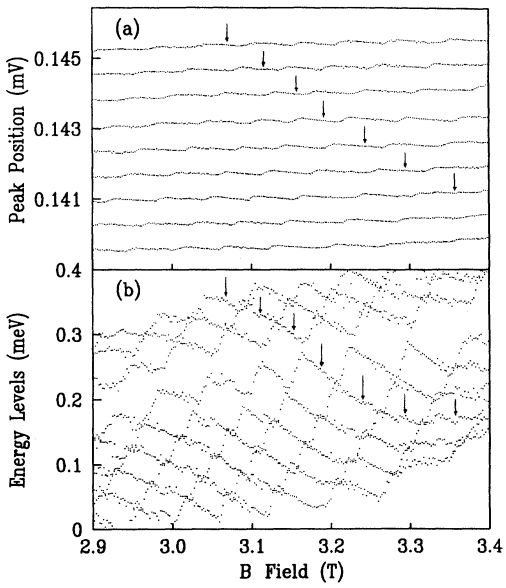

FIG. 17. (a) Peak position vs B for a series of consecutive conductance peaks. The arrow follows a particular state in the first Landau level as it moves through successive peaks. (b) Single-particle energy-level spectrum inferred from (a) as described in the text and McEuen *et al.* (1991). The zero of the energy scale is arbitrary.

voltage of the peaks was supposed to be simply e^2/C plus ΔE_N, apart from a factor e. This means that if we shift the Nth level by $-Ne^2/C$ we should have the spectrum of the ΔE_N. Figure 17(a) shows the positions in gate voltage of nine peaks as a function of B. In Fig. 17(b) we show the same data but after subtracting a constant difference voltage between successive peaks. The conversion from voltage to energy is not quite as straightforward as one might guess, and the details of how its done may be found in McEuan *et al.* (1991).

The spectrum we extract in this way is remarkably similar to that predicted in Fig. 15. It turns out that this simple model is not adequate to quantitatively predict the spectrum of Fig. 17(b). A more sophisticated treatment, however, analogous to the Hartree approximation, is very successful (McEuen *et al.*, 1992). It is thus clear that quantum mechanics is crucial at these energy scales, corresponding to 1 K or less. It is also clear that the suppression of charge fluctuations makes it possible to use the conductance as a spectroscopy of the energy levels of the Coulomb island.

VI. CONCLUSIONS

There still are many mysteries, some of which are being solved as I write this colloquium. A fascinating question is why the simple subtraction of the constant Coulomb term works so well. When we extract the level spectrum by subtracting Ne^2/C from the gate voltage of the Nth peak, we are tacitly assuming that the interactions among the electrons on the isolated region can be accounted for with the simple capacitance model. That is obviously not the case for atoms, so why does it appear to work here? Our transistors have only about 100 electrons on them in the gate voltage range we study, so they are like large atoms in terms of the number of electrons. In fact, the assumption of a constant Coulomb energy is not generally valid, and a self-consistent calculation shows that the Coulomb energy depends on the occupancy of the Landau levels (McEuen *et al.*, 1992).

The widths of the conductance peaks are, at present, limited by kT. We shall learn more about our single-electron transistors when we reach lower temperatures or make the tunneling matrix elements larger, so that we can measure the natural line shapes. Meir, Wingreen, and Lee have pointed out that our small transistor coupled to its leads is analogous to a transition-element impurity, with its localized d electrons, coupled to the s electrons in a host metal. When the average number of electrons on the Coulomb island is an integer, charge fluctuations are completely suppressed. This is analogous to the Kondo problem, in which the charge on the transition-element atom cannot fluctuate because of the large Coulomb energy required to add an electron, and only spin fluctuations occur. On the other hand, when the average number of electrons on the transistor is an integer plus $\frac{1}{2}$, the charge fluctuates, but by only $\pm\frac{1}{2}$. This is the case of the valence fluctuation problem in which the valence of the transition element fluctuates between two values, v and $v+1$. Line shapes for these cases can be predicted using what we know about these analogous problems.

Of course, as our accidental discovery of the conductance oscillations shows, it is the unexpected phenomena that are likely to be most exciting. I suspect we shall continue to be surprised for some time to come.

A number of clever applications of single-electron transistors have been proposed. Radio-frequency turnstile devices have already been demonstrated (Kouwenhoven *et al.*, 1991, 1992). By turning the two tunnel barriers of a single-electron transistor on and off sequentially, one generates a current $I=ef$ where f is the radio frequency. The intrinsic frequency response of single-electron transistors is very high because of their very small capacitance; for example, the capacitance of transistors like those whose characteristics are given in Fig. 9 is $\sim 10^{-16}$ F. However, despite their elegance, such turnstile devices are not expected to be very important technologically. They might eventually provide current standards, but their precision is, at this writing, still much lower than competing standards.

Likharev (1988) has proposed applications, such as multilevel logic, involving discrete devices which may become important if higher-temperature operation can be achieved. However, our guess is that, if single-electron transistors are to have a significant technological impact, it will be in much more unusual systems. It is likely that they will not be used as discrete components but rather in

internally coupled arrays.

As with most scientific discoveries, it is impossible to predict what the technological impact of the single-electron transistor will eventually be. The histories of the transistor and of the diode laser are good examples. Nowadays, new electronic technologies must compete with the industry based on silicon with its enormous past investment. Few new technologies are likely to compete successfully with conventional silicon technology. However, arrays of devices with novel properties may turn out to be valuable.

ACKNOWLEDGMENTS

Working with very small semiconductor structures is technologically very challenging. To do forefront research one needs the very best materials, the very best electron-beam lithography, and careful measurements at very low temperatures, as well as the insights of one's theoretical colleagues. The work I have described would not have come about without the group of great collaborators cited in footnote 1, who brought a diverse set of talents to bear on this exciting problem. Keeping such a strong collaboration together also requires the steadfast support of the Federal Government. I am grateful to acknowledge support from the National Science Foundation (Grant No. ECS-8813250) and by the United States Joint Services Electronics Program (under Contract No. DAAL03-89-C-0001).

REFERENCES

Altshuler, B. L., and P. A. Lee, 1988, Physics Today **41** (12), 37.

Amman, M., K. Mullen, and E. Ben-Jacob, 1989, J. Appl. Phys. **65**, 339.

Averin, D. V., and K. K. Likharev, 1991, in *Mesoscopic Phenomena in Solids,* edited by B. L. Altshuler, P. A. Lee, and R. A. Webb (Elsevier, Amsterdam), p. 173.

Averin, D. V., and Yu. V. Nazorov, 1990, Phys. Rev. Lett. **65**, 2446.

Beenakker, C. W. J., 1991, Phys. Rev. B **44**, 1646.

Beenakker, C. W. J., H. van Houten, and A. A. M. Staring, 1991, in *Single Charge Tunneling,* H. Grabert and M. H. Devoret, eds., NATO ASI Series B (Plenum, New York).

Darwin, C. G., 1930, Proc. Cambridge Philos. Soc. **27**, 86.

Field, Stuart, B., M. A. Kastner, U. Meirav, J. H. F. Scott-Thomas, D. A. Antoniadis, H. I. Smith, and S. J. Wind, 1990, Phys. Rev. B **42**, 3523.

Fock, V., 1928, Z. Phys. **47**, 446.

Glazman, L. I., and R. I. Shekhter, 1989, J. Phys. Conden. Matter **1**, 5811.

Kastner, M. A., R. F. Kwasnick, J. C. Licini, and D. J. Bishop, 1987, Phys. Rev. B **36**, 8015.

Kouwenhoven, L. P., A. T. Johnson, N. C. van der Vaart, C. J. P. M. Harmans, and C. T. Foxon, 1991, Phys. Rev. Lett. **67**, 1626.

Kouwenhoven, L. P., N. C. van der Vaart, A. T. Johnson, W. Kool, C. J. P. M. Harmans, J. G. Williamson, A. A. M. Staring, and C. T. Foxon, 1991, Z. Phys. B **85**, 367.

Kulik, I. O., and R. I. Shekhter, 1975, Zh. Eksp. Teor. Fiz. [Sov. Phys. JETP **68**, 623].

Kumar, A., S. E. Laux, and Frank Stern, 1989, Appl. Phys. Lett. **54**, 1270.

Kumar, A., S. E. Laux, and Frank Stern, 1990, Phys. Rev. B **42**, 5166.

Laux, S. E. , D. J. Frank, and Frank Stern, 1988, Surf. Sci. **96**, 101.

Likharev, K. K., 1988, IBM J. Res. Dev. **32**, 144.

McEuen, P. L., E. B. Foxman, J. Kinaret, U. Meirav, M. A. Kastner, N. S. Wingreen, and S. J. Wind, 1992, Phys. Rev. B (in press).

McEuen, P. L., E. B. Foxman, U. Meirav, M. A. Kastner, Yigal Meir, Ned S. Wingreen, and S. J. Wind, 1991, Phys. Rev. Lett. **66**, 1926.

Meir, Y., N. S. Wingreen, and P. A. Lee, 1991, Phys. Rev. Lett. **66**, 3048.

Meirav, U., M. Heiblum, and Frank Stern, 1988, Appl. Phys. Lett. **54**, 268.

Meirav, U., M. A. Kastner, H. Heiblum, and S. J. Wind, 1989, Phys. Rev. B **40**, 5871.

Meirav, U., M. A. Kastner, and S. J. Wind, 1990, Phys. Rev. Lett. **65**, 771.

Meirav, U., P. L. McEuen, M. A. Kastner, E. B. Foxman, A. Kumar, and S. J. Wind, 1991, Z. Phys. B **85**, 357.

Scott-Thomas, J. H. F., and Stuart B. Field, M. A. Kastner, H. I. Smith, and D. A. Antoniadis, 1989, Phys. Rev. Lett. **62**, 583.

Sivan, U., and Y. Imry, 1988, Phys. Rev. Lett. **61**, 1001.

Thouless, D. J., 1977, Phys. Rev. Lett. **39**, 1167.

van Houten, H., and C. W. J. Beenakker, 1989, Phys. Rev. Lett. **63**, 1893.

Wind, S. J., U. Meirav, and M. A. Kastner, 1990, in Microelectronic Eng. **13**, 411.

Zeller, H. R., and I. Giaver, 1969, Phys. Rev. **181**, 789.

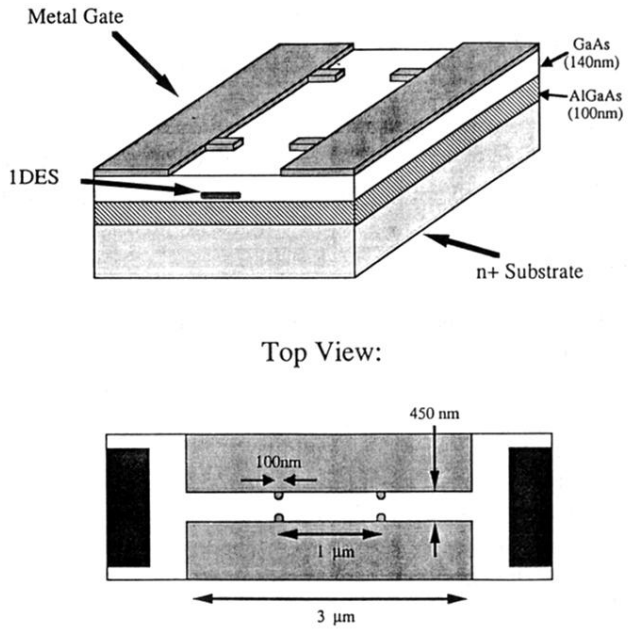

FIG. 7. Schematic drawings of device structure. Top: A one-dimensional electron gas (1DES) or narrow two-dimensional gas forms at the *top* GaAs-AlGaAs interface, with a density controlled by the substrate voltage V_g. Bottom: A top view showing typical dimensions of the top metal gate structure, which defines a narrow channel with two constrictions.