Measurements of the Surface Quantization in Silicon n- and p-Type Inversion Layers at Temperatures above 25 K

J. A. Pals
Philips Research Laboratories, Eindhoven, Netherlands
(Received 14 August 1972)

For an inverted semiconductor surface with a channel of minority carriers at the surface we show that, at sufficiently low temperatures, the Fermi-level position at the surface differentiated with respect to the number of minority carriers in the inversion layer has a characteristic minimum due to quantization of the motion perpendicular to the surface of the charge carriers in the inversion layer. This minimum can be experimentally determined by measuring the gate-bulk capacitance of a metal-oxide-semiconductor transistor (MOST). Experiments with silicon n- and p-type inversion layers demonstrate the quantization in the temperature range of about 25–75 K. Moreover, the effect of charged interface centers in the MOS system is investigated. These charges cause a broadening of the quantized energy levels for the motion perpendicular to the surface.

I. INTRODUCTION

The application of a sufficiently strong electric field of the appropriate polarity normal to the surface of a semiconductor causes the surface region to become inverted. A channel of minority carriers is formed, which is separated from the bulk by a depletion layer. The minority carriers in the inversion layer are bound to the surface in the potential well caused by the applied electric field. The motion of these carriers perpendicular to the surface must therefore be quantized, as Schrieffer first indicated.[1] Each energy level for the motion perpendicular to the surface brings about a two-dimensional subband, as the motion parallel to the surface of the charge carriers in the inversion layer is free. The quantization problem has been investigated by using the effective-mass approximation for the charge carriers in the inversion layer. The interaction between the carriers has been taken into account by determining a self-consistent solution of the coupled Schrödinger and Poisson equations. Numerical as well as approximate analytical solutions of the quantization problem have been given.[2–4]

A number of experiments have been performed at temperatures below 4.2 K which demonstrated very directly the quantization of the motion perpendicular to the surface for n-type inversion layers of silicon and indium antimonide.[5–9] For temperatures above 4.2 K up to room temperature, there is a substantial amount of experimental work pointing out the importance of quantization.[10–13] All these results, however, do not directly show unequivocally the existence of quantized energy levels for the motion perpendicular to the surface in this temperature range.

It is the purpose of this paper to report new experimental results for n- and p-type inversion layers at silicon surfaces which prove directly the quantization of the motion perpendicular to the surface in a temperature range of about 25–75 K.

In Sec. II it is shown that at a sufficiently low temperature the Fermi-level position at the surface differentiated with respect to the inversion-layer charge density shows a characteristic minimum as a function of the charge density due to the quantization of the motion perpendicular to the surface.

In Sec. III it is shown that this characteristic function can be experimentally determined by measuring the capacitance between gate and bulk contact as a function of the applied dc bias on the gate of a metal-oxide-semiconductor transistor (MOST). For a number of samples the experimental results are compared with the theoretical results.

The influence of localized charges in the semiconductor surface is investigated in Sec. IV, while the paper is concluded with a discussion of the results in Sec. V.

II. THEORY

To formulate the equations for charge carriers in an inversion layer with a quantized motion perpendicular to the surface, we shall only describe the case of electrons in an inversion layer at the surface of a p-type semiconductor. The equations for holes in an inversion layer at the surface of an n-type semiconductor are analogous.

The band bending for an n-type inversion layer is sketched in Fig. 1. The energy of the conduction band edge E_c is given as a function of the distance z from the surface and normal to the surface of the semiconductor. E_c is supposed to be independent of the coordinates parallel to the surface. The local potential variations owing to the presence of fixed charges in the surface are therefore neglected.

For the description of the motion of the electrons we shall use the effective-mass approximation. Moreover, we suppose that the temperature is

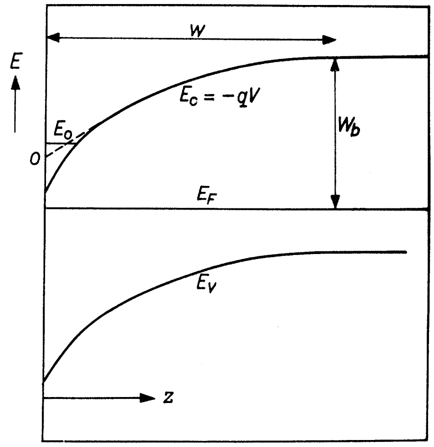

FIG. 1. Band bending at the surface of a p-type semiconductor with an n-type inversion layer.

sufficiently low and the field strength at the surface is sufficiently high as to be in the electric quantum limit, where only the lowest-energy level for the motion perpendicular to the surface is occupied.[2] For a multivalley conduction band we have then only to consider those valleys with the highest value of m_z, the effective mass perpendicular to the surface, since these valleys give the lowest-energy levels. For a silicon (100) n-type inversion layer these are the two valleys with $m_z = 0.98 m_e$; for a p-type silicon inversion layer only the heavy holes with $m_z = 0.5 m_e$ are considered. The lowest-energy level E_0 for the motion perpendicular to the surface with the corresponding normalized wave function $\Psi_0(z)$ and the potential $V(z)$ is determined by the coupled Schrödinger and Poisson equations.[2] The total number N_0 of electrons per unit square in the inversion layer on the energy level E_0 is given by

$$N_0 = \frac{n_v m_d kT}{\pi \hbar^2} \ln\left(1 + \exp \frac{E_F - E_0}{kT}\right) , \tag{1}$$

where m_d is the density-of-states mass for the motion parallel to the surface for the n_v equivalent valleys of the conduction band with the highest value of m_z.

By making use of the so-called depletion-layer approximation, the integration of the Poisson equation yields:

$$V(z) = -\frac{qN_a}{\epsilon_s}(wz - \tfrac{1}{2}z^2) + \frac{q}{\epsilon_s} N_0 \times \int_z^w (z' - z) |\Psi_0(z')|^2 dz' , \tag{2}$$

with a depletion-layer width w given by

$$w = \frac{1}{q}\left(\frac{2\epsilon_s}{N_a}(E_F + W_b)\right)^{1/2} . \tag{3}$$

The energy difference W_b between the conduction-band edge E_c and the Fermi level E_F in the bulk of the semiconductor is determined by the dopant concentration and the temperature.

The zero level of the potential is suitably chosen for our purposes to be such that if the electron concentration in the Poisson equation were neglected the potential at the surface would be equal to zero (Fig. 1). The first term in Eq. (2) is zero at $z = 0$. The position of the Fermi level E_F is measured with respect to this zero level.

A self-consistent solution of the equations for a quantized inversion layer can be obtained.[4] We have shown elsewhere,[14] that with the introduction of dimensionless quantities a general solution of the electric quantum limit problem can be given. Here we only need the dependence of the energy level E_0 on the total number of electrons N_0 in the inversion layer. We introduce a dimensionless energy level ϵ_0 and Fermi level ϵ_F by

$$\epsilon_0 = E_0\left(\frac{m_z}{q^2\hbar^2F^2}\right)^{1/3} , \quad \epsilon_F = E_F\left(\frac{m_z}{q^2\hbar^2F^2}\right)^{1/3} , \tag{4}$$

where F is the electric-field strength at the surface owing to the ionized impurities N_a only:

$$F = \frac{qN_a w}{\epsilon_s} . \tag{5}$$

In practice F can be considered to be a constant, independent of N_0, since the depletion-layer width w is practically constant for an inverted silicon surface. The total electron concentration per unit square, N_0, is normalized to the total number of acceptors in the depletion layer:

$$\alpha = \frac{N_0}{N_a w} = \frac{qN_0}{\epsilon_s F} . \tag{6}$$

With a variational calculus[14] we find for the normalized energy level ϵ_0, measured with respect to the zero level illustrated in Fig. 1, the following dependence on the normalized inversion-layer charge density α:

$$\epsilon_0 = \left(\frac{3}{2}\right)^{5/3} \frac{1 - \frac{3}{32}\alpha}{(1 + \frac{11}{32}\alpha)^{1/3}} . \tag{7}$$

By writing E_F as a function of E_0 and N_0 through inversion of Eq. (1) and by differentiating the normalized Fermi level ϵ_F with respect to the normalized inversion layer charge α, we obtain

$$\frac{d\epsilon_F}{d\alpha} = \left(\frac{m_z F}{q^2\hbar^2}\right)^{1/3} \frac{\epsilon_s}{q} \frac{dE_F}{dN_0} = \frac{d\epsilon_0}{d\alpha} + \gamma \frac{1}{1 - e^{-\alpha/\beta}} . \tag{8}$$

The dimensionless parameter γ and the normalized temperature β are given by

$$\gamma = \epsilon_s \left(\frac{m_z F}{q^2 \hbar^2} \right)^{1/3} \frac{\pi \hbar^2}{q n_v m_d} \ ,$$
$$\beta = \frac{q n_v m_d k T}{\pi \hbar^2 \epsilon_s F} \tag{9}$$

In Fig. 2 some results for $(d\epsilon_0/d\alpha)+\gamma$ which follow from Eq. (7) and $d\epsilon_F/d\alpha$ for a number of different temperatures are given for practical values of β and γ which are indicated in the figure. We see that for a sufficiently low temperature a characteristic minimum in the $d\epsilon_F/d\alpha$ curve occurs which is a direct consequence of the quantization of the motion perpendicular to the surface. The dashed curve in Fig. 2 represents the result of a conventional continuum calculation of $d\epsilon_F/d\alpha$. For this calculation we used the same parameters as above, but neglected the quantization in the inversion layer. We see that the minimum in the curve then disappears.

III. EXPERIMENTAL METHOD AND RESULTS OF MEASUREMENTS

The experiments were performed with common metal-oxide-semiconductor transistors (MOST's). A schematical cross section of such a device with an n-type inversion channel is given in Fig. 3. The diffused n^+ regions (source and drain) are used for making electrical contact with the inversion layer, which exists at the surface if a sufficiently high positive gate voltage V_g is applied to the gate metal which just overlaps the source and drain

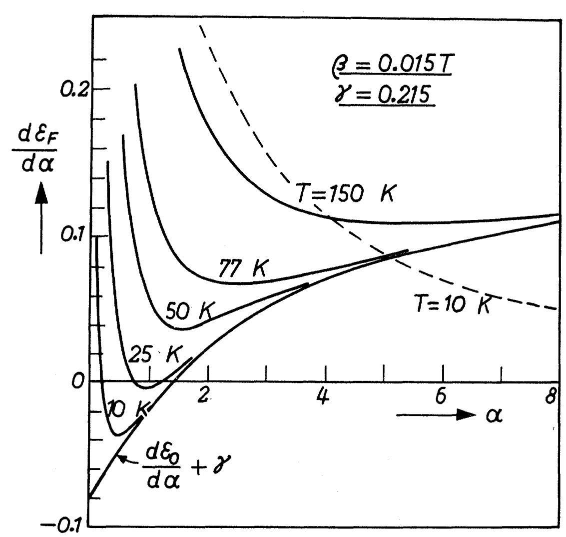

FIG. 2. Calculated curves of the normalized derivatives of the lowest energy level ϵ_0 and the Fermi level ϵ_F with respect to the normalized inversion layer charge α as a function of α at different temperatures. The dashed curve gives the result of a conventional continuum calculation, for which the quantization in the inversion layer was neglected.

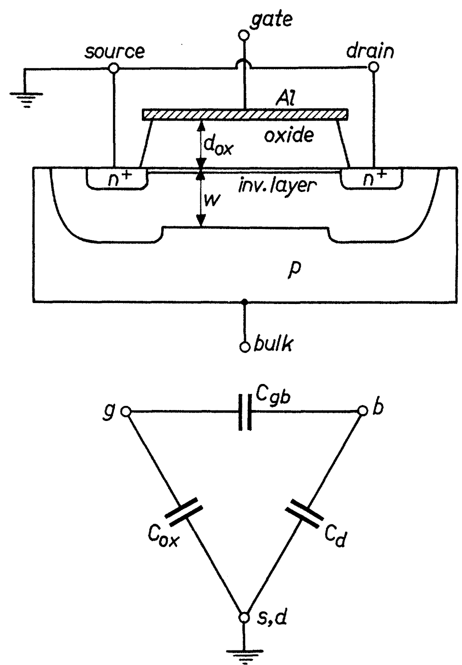

FIG. 3. Cross section of an n-channel MOST with the corresponding equivalent circuit.

regions. For our experiments the source and drain contacts are connected together. Our actual devices had a circular geometry. The channel length, the distance from source to drain diffusions, was 22 μm, and the total inversion layer area was 2.33×10^{-8} m^{-2}. The devices were made with standard silicon metal-oxide-semiconductor (MOS) technology, and the thickness of the thermally grown oxide was 1200 Å.

If an inversion layer exists, the device can be represented by the equivalent circuit also given in Fig. 3. All resistance elements, like the inversion-layer resistance and the bulk resistance, are neglected. We shall see later on in this section that this is allowed for the measurements. C_{ox} is the total oxide capacitance between the gate metal and the underlying source, drain, and inversion-layer regions. C_d is the total depletion-layer capacitance between the bulk of the semiconductor and the source, drain, and inversion-layer regions. The capacitance C_{gb} is very small, since the inversion layer forms a shield between the metal gate and the bulk of the semiconductor. A small change dV_g in the gate voltage with respect to the source, drain, and bulk contacts causes a small change in the inversion-layer charge density dN_0 and a small change in the depletion-layer width dw. A change in the depletion-layer width means that a certain amount of charge dQ_b flows through the bulk contact. For the capacitance C_{gb} it is found that if the surface area of the inversion layer is S,

$$C_{gb}=\frac{dQ_b}{dV_g}=qN_aS\,\frac{dw}{dV_g}=qN_aS\,\frac{dw}{dE_F}\,\frac{dE_F}{dN_0}\,\frac{dN_0}{dV_g}\quad . \tag{10}$$

$q\,dN_0/dV_g$ is practically equal to the oxide capacitance per unit square, ϵ_{ox}/d_{ox}, as the depletion-layer charge is almost independent of the gate voltage for an inverted silicon surface. With the use of Eq. (3), we obtain

$$C_{gb}=\frac{\epsilon_{ox}\epsilon_s S}{d_{ox}\,wq^2}\,\frac{dE_F}{dN_0}\quad . \tag{11}$$

We see that the capacitance C_{gb} is proportional to the quantity dE_F/dN_0, which, as shown in Sec. II, has a characteristic minimum as a function of N_0 due to the quantization in the inversion layer .

Our measurements consist of determining the voltage-transfer ratio from gate to bulk. This ratio is equal to the ratio of C_{gb} to the much larger value of C_d, which is for our devices typically 40 pF. A small-signal ac voltage of 100 mV at a frequency of 20 kHz is applied to the gate contact. The in-phase component of the resulting ac voltage on the bulk contact is measured with a lock-in amplifier. By varying the dc gate voltage V_g, the output of the lock-in amplifier is recorded on an x-y recorder as a function of V_g.

For the measuring frequency the approximation of neglecting the resistance elements is allowed. For the n-channel device reported in Fig. 4(a) we have, for instance, for $\alpha=1$, which corresponds to $N_0=3.75\times10^{15}$ m^{-2}, a source-drain resistance $R=340\ \Omega$ giving a ωRC_{ox} product of 3×10^{-4}, and the neglection of the resistance in the R-C transmission line between gate- and source-bulk contacts is allowed. The bulk resistance can be neglected because of the much higher input impedance of the lock-in amplifier. We have also experimentally checked this point by varying the measuring frequency from 1 to 500 kHz without finding any influence on the recorded curves.

As there is a parasitic header capacitance of about 0.2 pF between the gate and bulk contact which is larger than the internal capacitance value C_{gb} (0.001–0.01 pF for our devices), this header capacitance has to be compensated for, and we can only determine the variations in C_{gb} and not its absolute value. From the measured variations in C_{gb} we determine the variations in $d\epsilon_F/d\alpha$; the proportionality factor is obtained by combining Eqs. (8) and (11). As a typical example, the device reported in Fig. 4(a) has the following numerical relation: $C_{gb}=0.07\,\text{pF}\ (d\epsilon_F/d\alpha)$. The dimensionless inversion-layer charge α is related to the applied voltage V_g by

$$\alpha=\frac{qN_0}{\epsilon_s F}=\frac{\epsilon_{ox}}{\epsilon_s F d_{ox}}\,(V_g-V_{th})\quad , \tag{12}$$

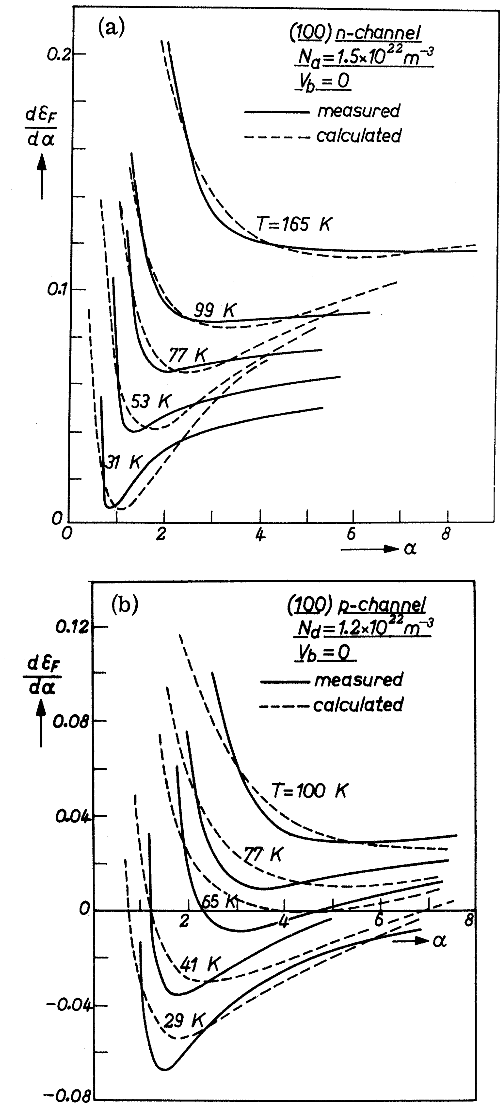

FIG. 4. Measured and calculated curves of $d\epsilon_F/d\alpha$ as a function of α at different temperatures, and with a bulk bias $V_b=0$: (a) for n-channel sample with (100) orientation; (b) for a p-channel sample with (100) orientation.

where V_{th} is the so-called threshold voltage of the MOST. Again we have made the approximation of a constant depletion-layer width with varying V_g for an inverted silicon surface.

In this way we performed measurements for a number of n- and p-channel MOST's, for which all parameters, such as dopant concentration in the bulk, threshold voltage, and oxide thickness, were determined separately. In Figs. 4(a) and 4(b) the results for an n-channel and p-channel silicon sample with a (100) surface orientation are given for different temperatures. The recorded curves of the ac bulk voltage as a function of the dc gate voltage are multiplied with the appropriate scaling factors to obtain $d\epsilon_F/d\alpha$ as a function of α. The vertical position of the measured curves is matched, as only the variations in C_{gb} can be measured.

As known the depletion-layer width w and the field strength F may be varied by applying a dc bulk voltage V_b between the bulk and source-drain contact of the MOST. The effect of the bulk bias V_b on the $d\epsilon_F/d\alpha$ curve as found by measurements and calculations, is given for the same samples in Figs. 5(a) and 5(b) for a temperature of $T = 38$ K.

The same measurements were also done with p-channel silicon samples with a (110)- and (111)-surface orientation. The results of a measurement for three samples with a different surface orientation at $T = 38$ K are given in Fig. 6. The only difference between these samples is the surface orientation. They are cut from the same silicon bar and made under the same technological circumstances. We see a flattened minimum in the $d\epsilon_F/d\alpha$ curves for the (110)- and (111)-surface oriented samples. This cannot be explained by a difference in effective mass values for the different surface orientations, as the effective mass is isotropic for holes in first approximation. A possible reason for the difference may be due to the charged centers in the interface for which the density is known to be larger for an oxidized silicon surface of (110) and (111) orientation compared with a (100)-oriented surface.[15,16] On the grounds of differences in the threshold voltage, our (110) and (111) samples have a number of charged interface centers per unit square of 6.5×10^{15} and 7.5×10^{15} m^{-2}, respectively, compared with 2×10^{15} m^{-2} of the (100)-oriented sample. In Sec. IV the effect of these charged interface centers on the measurements shall be investigated further.

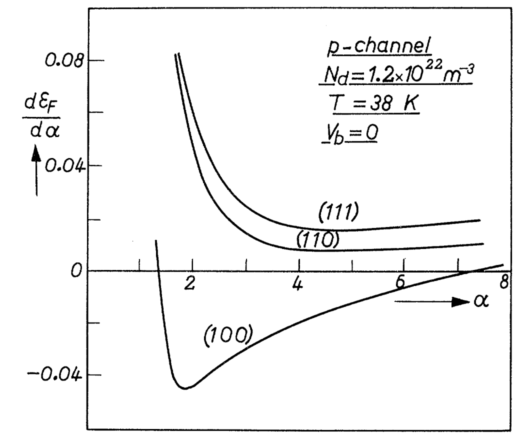

FIG. 6. Measured curves of $d\epsilon_F/d\alpha$ as a function of α at $T = 38$ K and with $V_b = 0$ for p-channel samples with a different surface orientation.

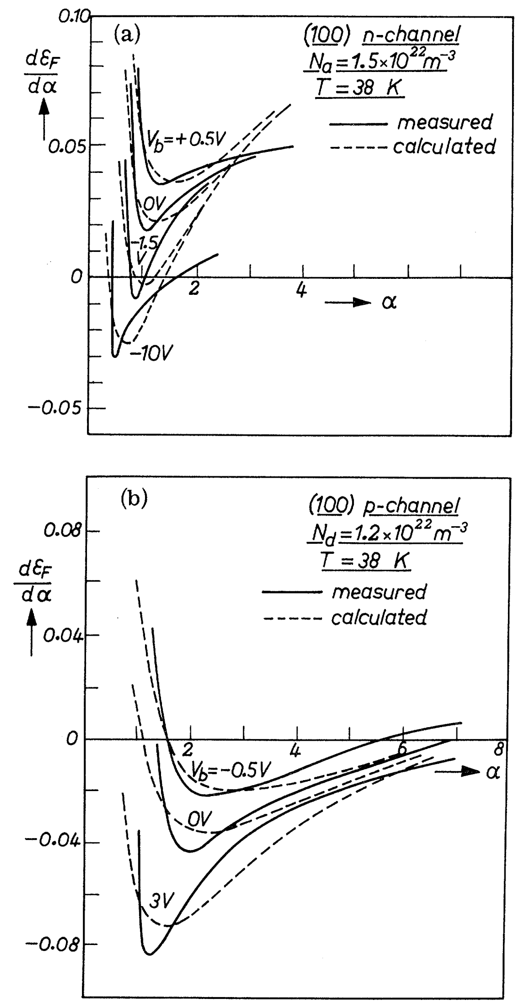

FIG. 5. Measured and calculated curves of $d\epsilon_F/d\alpha$ as a function of α at $T = 38$ K and with different applied bulk biases V_b: (a) for an n-channel sample with (100) orientation; (b) for a p-channel sample with (100) orientation.

IV. EFFECT OF CHARGED INTERFACE CENTERS

To investigate whether interface charges cause a flattening of the minimum in the $d\epsilon_F/d\alpha$ curve, we performed the following experiment: a p-channel sample with a (100)-surface orientation was first measured at 38 K. After this measurement the sample was exposed to x rays from an x-ray tube with a tungsten anode and with an anode voltage of 150 kV. The intensity of the irradiation was 10^4 R/min. After a given exposure time the sample was again measured, under the same experimental circumstances as before irradiation. In Fig. 7 the resulting measured curves are given with the exposure time as a parameter. The den-

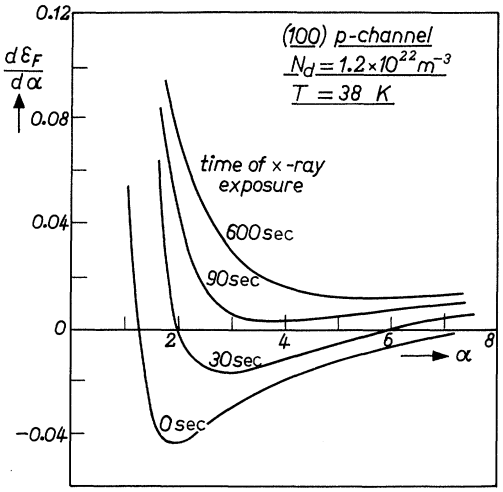

FIG. 7. Measured curves of $d\epsilon_F/d\alpha$ as a function of α at $T = 38$ K and with $V_b = 0$ for a p-channel (100) oriented sample after different times of x-ray irradiation.

sity of extra-charged interface centers caused by the x-ray irradiation[17] was determined by the change in the threshold voltage. For an exposure time of 0, 30, 90, and 600 sec we found values of 2×10^{15}, 5.5×10^{15}, 8.5×10^{15}, and 1.7×10^{16} m^{-2} interface charges, respectively. We observe that the increase in density of interface charges indeed causes a strong flattening of the minimum in the measured $d\epsilon_F/d\alpha$ curves. The density of extra charged interface centers in this experiment, which causes the flattening in the minimum, is of the same order as the extra density of interface charges of the (110) and (111)-oriented samples. compared with the (100)-oriented sample of Sec. III F. This experiment therefore gives a strong indication that the different results for the (110)- and (111)-oriented samples can be explained by the higher density of charged interface centers.

A theoretical calculation of the effect of fixed interface charges on the quantized inversion layer is very difficult. A charged center at the surface causes a local variation in the potential both perpendicular and parallel to the surface, and the Schrödinger equation can no longer be separated in a free motion parallel to the surface and a quantized motion perpendicular to the surface. The problem is further complicated by the screening of a charged center by a redistribution of the free carriers in the inversion layer. Several approximate solutions of this problem are given.[2,3]

A random distribution of fixed charges results in in the energy level E_0 for the motion perpendicular to the surface being no longer a distinct level but becoming a broadened level. Instead of calculating this effect, we introduce an *ad hoc* density of states with a broadened energy level by

$$D(E) = \frac{n_v m_d}{\pi \hbar^2}\, \frac{1}{1+\exp(E_0 - E)/E_s}\ , \tag{13}$$

in which E_s is a measure for the broadening of the energy level E_0. For $E_s \to 0$ we reach the situation with a distinct energy level E_0. The total inversion-layer charge density for the electric quantum limit is now given by

$$N_0 = \int_{-\infty}^{\infty} \frac{D(E)}{1+\exp(E-E_F)/kT}\, dE\ . \tag{14}$$

Supposing that the energy level E_0 depends in the same way on N_0 as in the case of a sharp level [Eq. (7)], we now straightforwardly calculate the $d\epsilon_F/d\alpha$ curve for different values of E_s. We have performed these calculations numerically for a silicon p-type inversion layer with (100)-surface orientation and the results are given in Fig. 8. With a spreading parameter value of E_s of about 5 meV, the difference in behavior can be explained between the (110), (111), and the x-ray-irradiated (100) samples on the one hand and the (100) oriented sample without an x-ray exposure on the other.

This value of 5 meV is quite reasonable. The calculations of Stern and Howard[2] for an n-type inversion layer and one positive elementary charge at the surface show the existence of a bound state of about 4 meV below E_0, if the effect of screening is taken into account. Owing to the interaction between the neighboring charged centers and owing to the spread in the distance of these centers from the oxide-semiconductor interface, we have a

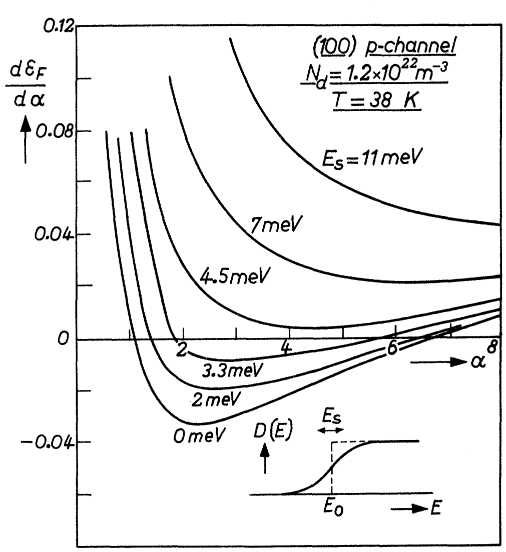

FIG. 8. Calculated curves of $d\epsilon_F/d\alpha$ as a function of α for a broadened lowest-energy level E_0; the broadening is characterized by the parameter E_s.

spread in possible energies around E_0 of the order of 4 meV. This also agrees with infrared measurements of Wheeler and Ralston[8] who find a width of the energy levels for the motion perpendicular to the surface of about 3 meV for their samples.

We therefore conclude that the spread of 5 meV which we need to explain the measurements for (110)-, (111)-, and (100)-oriented samples after x-ray irradiation agrees with the spread in energy which is expected on the grounds of theoretical and experimental results obtained earlier .

V. DISCUSSION

For all the samples with a (100)-surface orientation and with p- or n-type inversion layers, we have found a qualitative agreement between measurements and calculations for the Fermi-level position at the surface differentiated with respect to the inversion-layer charge density. A characteristic minimum in this quantity as a function of the inversion-layer charge density occurs in a temperature range of about 25–75 K. At higher temperatures the minimum disappears due to thermal broadening. At lower temperatures the measurement is no longer possible because of the freezing out of the majority carriers in the bulk of the semiconductor.

The much flatter minimum in the curves measured for the (110)- and (111)-oriented p-channel samples can be ascribed to the higher density of interface charges for these samples, compared with the (100)-oriented samples. This is experimentally demonstrated by introducing extra charged surface centers in (100)-oriented samples with x-ray irradiation. A broadening of the energy level E_0 for the motion perpendicular to the surface explains the flattening of the minimum in the measured curves.

Although there is a rather good qualitative agreement, some systematic deviations between experimental and calculated curves still exist. For instance, we see that the measured curves show a sharper minimum than the calculated curves. This may be due to the simplifying assumptions which are made for the calculations.

(i) We have used the effective mass approximation to describe the motion perpendicular to the surface. This is questionable[2] in view of the very small thickness of the inversion layer, which is of the order of 100 Å, and of the strong electric field perpendicular to the surface in the inversion layer.

(ii) The interaction between the carriers in the inversion layer is approximated by taking into account the average space-charge density of the carriers for the solution of the Poisson equation, as is done in the Hartree approximation for the solution of the wave functions for atoms. Alferieff and Duke[18] have indicated that the density of states may be more discontinuous than we find with our approximation, and this may contribute to the sharper minimum in the measured curves.

To conclude we remark that clear evidence of surface quantization in inversion layers existed at temperatures below 4.2 K. On the other hand, at higher temperatures previous experimental evidence was not as unequivocal. This paper, however, presents new and direct evidence for the quantization in p- and n-type silicon inversion layers in a temperature range of about 25–75 K.

[1]J. R. Schrieffer, in *Semiconductor Surface Physics*, edited by R. H. Kingston (University of Pennsylvania Press, Philadelphia, 1957), p. 55.
[2]F. Stern and W. E. Howard, Phys. Rev. **163**, 816 (1967).
[3]E. D. Siggia and P. C. Kwok, Phys. Rev. B **2**, 1024 (1970).
[4]F. Stern, J. Comput. Phys. **6**, 56 (1970).
[5]A. B. Fowler, F. F. Fang, W. E. Howard, and P. J. Stiles, Phys. Rev. Lett. **16**, 901 (1966).
[6]F. F. Fang and P. J. Stiles, Phys. Rev. **174**, 823 (1968).
[7]M. Kaplit and J. N. Zemel, Phys. Rev. Lett. **21**, 212 (1968).
[8]R. G. Wheeler and R. W. Ralston, Phys. Rev. Lett. **27**, 925 (1971).
[9]Y. Katayama, N. Kotera, and F. Komatsubara, Jap. J. Soc. Appl. Phys. Suppl. **40**, 214 (1971).
[10]T. Sato, Y. Takeishi, H. Hara, and Y. Okamoto, Phys. Rev. B **4**, 1950 (1971).
[11]G. Dorda, J. Appl. Phys. **42**, 2053 (1971).
[12]S. Tansal, A. B. Fowler, and R. F. Cotellessa, Phys. Rev. **178**, 1326 (1969).
[13]J. A. Pals, Phys. Rev. B **5**, 4208 (1972).
[14]J. A. Pals, Phys. Lett. A **39**, 101 (1972).
[15]P. V. Gray and D. M. Brown, Appl. Phys. Lett. **8**, 31 (1966).
[16]G. Abowitz, E. Arnold, and J. Ladell, Phys. Rev. Lett. **18**, 543 (1967).
[17]E. Kooi, Philips Res. Rep. **20**, 595 (1965).
[18]M. E. Alferieff and C. B. Duke, Phys. Rev. **168**, 832 (1968).