Transport at low electron density in the two-dimensional electron gas of silicon MOSFETs
Heemskerk, Richard

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
1998

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 2

Experimental system: silicon field effect transistors

2.1 Introduction

The MOSFET (an acronym for Metal Oxide Semiconductor Field Effect Transistor) has been developed as an electronic switch to regulate the conduction between two terminals by a third terminal, the gate. The change from conduction to insulation is studied here under uniquely tailored circumstances to get as closely as possible to the underlying physical processes. The conduction of the electrons occurs at the interface of the semiconductor silicon (Si) and the insulating silicdioxide (SiO$_2$) (Fig. 2.1). The electrons at the interface form a conducting sheet, which at low temperatures, for example at the boiling point of liquid helium $T \leq 4.2K$, means that they are trapped in a triangular quantum well perpendicular to the interface. Hence, electrons can only move along the interface, and form a 2-dimensional electron gas (2DEG). In this thesis we study the change of conduction of these 2DEG electrons from a conducting state to an insulating state as a function of electron density.

The electron density is controlled by the MOS-capacitor (Sec. 2.2), where the band diagram and the threshold voltage is discussed and is corrected for low temperatures. The specific requirements for the present experiments are discussed in (Sec. 2.3) as well as the technique to determine the carrier density accurately.
Chapter 2. Experimental system: silicon field effect transistors

Figure 2.1: Schematic drawing of a MOSFET with three relevant terminals: source, drain, and gate

2.2 Control of carrier density

2.2.1 Metal-oxide-silicon capacitor

The metal-oxide-silicon (MOS) capacitor consists (Fig. 2.2) of three stacked layers: a metal (M), an oxide (O), and a semiconductor (S) layer. The semiconductor is usually silicon and the oxide silicon dioxide. The oxide is a thin layer of about 100nm thick. The used silicon substrates are 5cm in diameter and about 275µm thick. The metal can in principle be any metal, but often aluminium is used, the thickness varies from 25nm for the MOSFETs for this thesis to 2µm thick, the more commonly used thickness for practical devices. With the MOS-capacitor we control the charge density near the interface of the oxide and the semiconductor, like in an ordinary capacitor. The charge at the interface plays the major role in the operation of a MOSFET.

The MOSFET has a similar construction, but has two additional contacts (terminals) to the silicon at the interface of the Si-SiO₂ (Fig. 2.1). The conduction between these two contacts to the silicon (called source and drain) can be regulated by the voltage applied to the gate, the third terminal. The sign of the required gate voltage for conduction depends on the type of silicon, i.e. whether we have n-type or p-type silicon. With p-type silicon the two contacts are n-doped, so there is a high electron density in the contacts
in contrast with the substrate which has a high hole density and a negligible electron concentration. With p-type silicon an n-channel MOSFET is formed, and the conducting state is reached by attracting electrons to the Si-SiO₂ interface with a positive voltage on the gate; there is no change in charge carrier in the transport of charge from one contact to the other. For lower voltages the density of the electrons is reduced at the interface, and the conduction is lower. For MOSFETs with n-type silicon the conduction is achieved by applying a low voltage (often a negative voltage), and electrons are driven from the interface into the bulk, while a high concentration of holes provides the conducting path between the two contacts with a high hole density. The present study focuses on the metal-insulator transition in a 2-dimensional electron gas, formed in p-type silicon MOSFETs. Hence the transition occurs as a result of the lowering of the gate voltage from a positive value.

### 2.2.2 The band diagram and the Fermi level

In a MOS-capacitor the position of the Fermi-energy is different between the conductor and the semiconductor (Fig. 2.2). The difference in Fermi-energy is referred to as the difference in workfunction. For the semiconductor the Fermi-energy is in the energy gap somewhere between the conduction and the valence band (Fig. 2.3), the Fermi-energy of a metal is well into the conduction band. The difference can also be expressed in the screening length, which is short in the conductor (atomic scale), whereas in the semiconductor it depends on the gate voltage and can be several microns long. This is due to the much lower concentration of charge carriers.

The Fermi-energy is determined by the hole and electron concentration
in the semiconductor. It is closer to the valence band energy \( E_v \) when there are more holes than electrons, while it is closer to the conduction band energy \( E_c \) in the opposite case. The electron density \( n(T) \) and the Fermi-energy \( E_F \) are related by:

\[
n(T) = \frac{1}{2\pi^2} \left( \frac{2m_e^*}{\hbar^2} \right)^{\frac{3}{2}} \int_0^\infty \frac{E_c^\frac{3}{2}dE}{1 + \exp \left[ \frac{E - E_F + E_c}{k_B T} \right]} \tag{2.1}
\]

and for the hole density \( p(T) \):

\[
p(T) = \frac{1}{2\pi^2} \left( \frac{2m_h^*}{\hbar^2} \right)^{\frac{3}{2}} \int_0^\infty \frac{E_v^\frac{3}{2}dE}{1 + \exp \left[ \frac{E - E_F + E_v}{k_B T} \right]} \tag{2.2}
\]

with \( k_B T \) the thermal energy, \( \hbar \) Planck’s constant[2, 3], \( E_c \) is the conduction band edge, and \( E_v \) is the valence band edge. Furthermore, \( m_e^* \) is the effective mass of the electrons in the conduction band and \( m_h^* \) is the effective mass of holes in the valence band.

The semiconductor is called non-degenerate, when the Fermi-energy is far away from the band edges:

\[
E_c - E_F \gg k_B T \tag{2.3}
\]

\[
E_F - E_v \gg k_B T \tag{2.4}
\]

Under these conditions the integrals of Eqs. 2.1 and 2.2 can be approximated
2.2. Control of carrier density

by a Maxwell-Boltzmann distribution for the electron and hole density:

\[
\begin{align*}
n(T) &= N_e(T) \exp \left[ \frac{E_F - E_c}{k_B T} \right] \\
p(T) &= P_v(T) \exp \left[ \frac{E_v - E_F}{k_B T} \right]
\end{align*}
\]  (2.5)

with \(N_e(T)\) and \(P_v(T)\) the effective electron and hole density of states at the band edge.

The expression relating the Fermi-energy to the hole and electron concentrations is general, and holds for an intrinsic semiconductor, for which \(p(T) = n(T)\) as in undoped silicon, as well as for extrinsic semiconductors, for which \(p(T) \neq n(T)\). The position of the Fermi-energy is changed by adding dopants to the semiconductor (donors or acceptors). The donors such as phosphorous (P) in Si will ionize by releasing an electron to accommodate the P in the diamond structure of silicon. The electron concentration is increased by these free electrons. Acceptors (e.g. boron(B) in Si) are dopants that ionize by extracting an electron from the valence band, the hole concentration is increased and the Fermi level in the bandgap is lowered. When no potential is applied to the semiconductor, the semiconductor is neutral and hence:

\[p(T) + N_{di} = n(T) + N_{ai}\]  (2.6)

with \(N_{di}\) the ionized donor concentration and \(N_{ai}\) the ionized acceptor concentration. This equation, together with Eq. 2.5 describes how the Fermi-energy changes when acceptors or donors are added to the semiconductor.

Until now the semiconductor was considered to be neutral. In reality for any application of a MOS structure a voltage is applied between the semiconductor and the gate (metal), which changes the charge density near the interface(Fig. 2.4). As an example p-type silicon (acceptor doped) is taken. For a specific value of the voltage, the flatband voltage, which equals the difference in workfunctions of the metal and the semiconductor, there is no charge build-up in the MOS-capacitor. The hole and electron concentration are the same throughout the whole semiconductor (Fig. 2.4a). When a lower voltage is applied holes are accumulated at the interface, while electrons are removed. The major contribution to the charge density will come from the accumulated hole concentration, because \(p(T) > n(T)\). Consequently, the Fermi-energy will also change compared to the band edges. The bands are curved as in Fig. 2.4b. For higher voltages than the flatband
voltage the bending will be in the opposite direction. The hole concentration decreases and the electron concentration increases. This phenomenon is called depletion as long as the hole concentration is larger than the electron concentration, but the major contribution to the charge in this situation comes from the ionized acceptors \( N_{ai} > p(T) > n(T) \), as in Fig. 2.4c. For even higher voltages inversion occurs, when the electron concentration becomes larger than the hole concentration, although it is still smaller than the acceptor concentration (Fig. 2.4d.). When the electron concentration is increased until the acceptor concentration is reached, then the applied voltage is the threshold voltage. Above the threshold voltage the enhanced electron concentration is the dominant contribution to the charge density at the interface in the semiconductor (strong inversion regime). When the electron density is further increased, the Fermi-energy at the interface might be in the conduction band, and the semiconductor is non-degenerate.

### 2.2.3 The threshold voltage

Theoretically the threshold voltage can be calculated from the bandbending in the semiconductor. At this voltage the bandbending is two times the energy difference between the Fermi-energy in the bulk and the intrinsic energy. The Fermi-energy in the bulk can be calculated with Eqs. 2.5 and 2.6, when the doping concentration is known. The intrinsic energy is derived from Eq. 2.5 by taking the hole density equal to the electron density.

The electrostatic potential needed to determine the bandbending at the threshold voltage can be found by solving Poisson’s equation:

\[
\frac{d^2 \phi(z)}{dz^2} = -\frac{\rho(z)}{\epsilon}
\]  

with \( \rho(z) \) the charge density, and \( \epsilon \) the dielectric constant. The charge density \( \rho(z) \) is dominated by the density of ionized impurities, while the contribution of the electron and hole density are neglected. This is typical for a semiconductor and in this case the semiconductor is said to be depleted. An effective depletion width \( W_{dep} \) can at \( T = 0 \) be defined for which:

\[
\rho(z) = \begin{cases} 
  N_{ai} & \text{if } 0 < z < W_{dep} \\
  0 & \text{if } z > W_{dep}
\end{cases}
\]

The depletion width can be calculated by integrating Eq. 2.7 and using the electrostatic potential for the bandbending \( (\Psi_{BB}) \). This results in:
2.2. Control of carrier density

Figure 2.4: The bandbending in p-type silicon when a. the flatband voltage is applied, b. a lower voltage is applied (accumulation), c. a higher voltage is applied (depletion), d. a much higher voltage is applied (inversion).
Chapter 2. Experimental system: silicon field effect transistors

\[ W_{dep} = \sqrt{\frac{2\varepsilon_S \Psi_{BB}}{eN_{ai}}} \]  

(2.9)

and the total charge in the semiconductor is:

\[ q_{dep} = eN_{ai}W_{dep} \]  

(2.10)

When the total charge in the semiconductor is known, then the threshold voltage \( V_{th} \) on the MOS-capacitor can be calculated from:

\[ V_{th} = V_{FB} + \frac{q_{dep}d_{ox}}{\varepsilon_{ox}} + \frac{2E_i - E_F(bulk)}{e} \]  

(2.11)

with \( V_{FB} \) the flatband voltage, the second term the voltage across the oxide and the last term the voltage across the silicon \( (\Psi_{BB}) \). The equation holds for p-type silicon, whereas for n-type silicon a minus sign appears in the second term.

2.2.4 Modifications at low temperatures

The inversion layer in the MOS-structure becomes a so called 2-dimensional electron gas, 2DEG, at low temperatures. It is a result from the fact that the electrons are trapped in the triangular potential well and form discrete energy levels with a mutual separation larger than the thermal energy. It is assumed that only the lowest subband is filled and that for the conduction only the two degrees of freedom are left for movement parallel to the interface. It is important to know the electron density of the 2DEG. At low temperatures some modifications in the relation of the Fermi-level with respect to the band diagram and the threshold voltage appear.

The MOS-structure is described as a capacitor with the inversion layer near the interface, and the electron density of the 2DEG is the charge per area on a capacitor with a capacitance equal to the oxide capacitance. The relevant voltage is the voltage across the capacitor relative to the threshold voltage. For a parallel-plate capacitor this is:

\[ n_s = \frac{c_{ox}}{e}(V_g - V_{th}) \]  

(2.12)

with \( n_s \) the electron density and \( c_{ox} \) the oxide capacitance, both per area.

Eq. 2.12 is an approximation. At the threshold voltage the change is not abrupt, because the change in the major charge density from ionized dopants...
2.2. Control of carrier density

to free electrons is gradual. There is still a small contribution to the charge from ionized dopants, which deplete just above the threshold voltage. The effect of this gradual change is clearly visible in Fig. 2.5 where the capacitance change is shown as a function of applied voltage at room temperature. At the threshold voltage it is strongly reduced, and only for voltages more than a 0.5 V larger than \( V_{th} \) the reduction is only a few percent.

A closer look at the validity of Eq. 2.12 is needed for low temperature operation, in particular the derivation of the threshold voltage and the electron density at the interface. The Fermi-energy in the bulk will decrease for p-type silicon, because the electron and hole density both decrease and the concentration of dopants becomes relatively larger. When the temperature is low enough the Fermi-energy will cross the ionization level of the acceptors, and part of the acceptors will be neutralized. The Fermi-energy is then between the valence band and the ionization level of the acceptor (\( E_a \)), which is 45meV for B in Si and 67meV for Al in Si. The ratio between the ionized (\( N_{ai} \)) and neutral (\( N_{an} \)) acceptor density is described by the Boltzmann-factor in a similar way as the electron density with respect to the effective electron density of states at the conduction band edge (\( N_c(T) \)):

\[
\frac{N_{ai}}{N_{an}} = \frac{1}{2} \exp \left[ \frac{E_F - E_a}{k_B T} \right]
\] (2.13)

The total charge in the bulk is zero. Therefore the electron density can be neglected and the hole density will be equal to the ionized acceptor density to maintain charge neutrality. This lead for Eq. 2.13 together with Eqs. 2.5 and 2.6 to the following expression for the hole density:

\[
p(T) = N_v \exp \left[ \frac{E_v - E_F}{k_B T} \right] = \frac{N_a}{1 + 2 \exp \left[ \frac{E_v - E_F}{E_F} \right]} \] (2.14)

with \( N_a \) the total acceptor density. This equation can be rewritten as an equation for the Fermi-energy:

\[
E_F = E_a - k_B T \ln \left( \frac{1}{4} \left[ \sqrt{1 + 8 \frac{N_a}{N_c} \exp \frac{E_a - E_v}{k_B T}} - 1 \right] \right)
\] (2.15)

which can be approximated for low temperatures by:

\[
E_F = \frac{E_a + E_v}{2} + \frac{k_B T}{2} \ln \left( \frac{N_a}{2N_c} \right)
\] (2.16)
As the temperature goes to zero all the states in the valence band will be occupied and none of the acceptors will be ionized. The Fermi-energy will be in the middle of the valence band and the ionization level of the acceptors (Eq. 2.16). At a temperature of 4.2 Kelvin the Fermi-energy deviates hardly from the ground state.

At low temperatures the Fermi-energy will change by bandbending at the interface, as a response to the voltage applied to the MOS-capacitor, compared to the flat band voltage. The flat band voltage itself is also different at low temperature, because it is determined by the difference in workfunction between the metal and the semiconductor. The Fermi-energy of the latter has changed strongly from somewhere in the middle of the gap to close to the valence band for p-type silicon.

In practice the Fermi-energy will be pinned at the ionization level of the acceptors, when there is also a small amount of donors in the silicon. Then the donors are ionized and the released electrons will occupy part of the acceptor states just above the valence band. For \( T = 0K \) the Fermilevel is at the ionization level. At low temperature just above \( T = 0K \) the Fermi-energy is not distinguishable from the ionization energy.

When the bandbending is sufficiently large that the Fermi-energy will cross the ionization level of the acceptors, the acceptors will be ionized and a potential for the bandbending will build up with the same rate as at room temperature. This continues until the threshold voltage is reached for which the electron density at the interface is equal to the ionized acceptor concentration. With Eq. 2.5 the corresponding Fermi-energy can be calculated, which will be very close to the conduction band due to the low temperatures. The total bandbending \( (E_{BB}^{max}) \) is given approximately by:

\[
E_{BB}^{max} = E_g - (E_a - E_v)
\]

with \( E_g \) the bandgap energy. The maximum depletion width is in this case:

\[
W_{dep}^{max} = \sqrt{\frac{2\epsilon_s E_{BB}^{max}}{e^2 N_{ai}}}
\]

and the total charge as well as the threshold voltage can be calculated as in Sec. 2.2.2.

The free electron density of the 2DEG at the threshold voltage can be determined with Eq. 2.5. The Fermi-energy at the interface is calculated by taking the electron density at the interface equal to the dopant concentration. With Eqs. 2.5 and 2.9, which are modified to include the \( z \)-dependence of...
the electron density, the Fermi-energy is found to change linearly from the value found at the interface to zero at $W_{dep}^{max}$. The result is:

$$n(z) = N_{ai} \exp \left[ -\frac{e^2 N_{ai} W_{dep}^{max} z}{\epsilon_{Si} k_B T} \right]$$  \hspace{0.5cm} (2.19)$$

This equation can be easily integrated from zero to $W_{dep}^{max}$ giving the electron density per area:

$$n_s = \frac{N_{ai} W_{dep} k_B T}{E_{BB}}$$  \hspace{0.5cm} (2.20)$$

with $n_s$ the sheet electron density of the 2DEG at the threshold voltage. Notice that the electron density is nonzero at the threshold voltage unlike the extrapolation to zero electron density often applied for a 2DEG (see Eq. 2.12). The electron density is still small compared to those used in practice. This will be demonstrated in Sec. 2.2.5

### 2.2.5 Determination of carrier density

Capacitance-voltage (CV) measurement of a MOS-capacitor or MOSFET are used to obtain information about the experimentally relevant quantities, the capacitance of the oxide $C_{ox}$, the thickness of the oxide $d_{ox}$, the flat band voltage $V_{FB}$, the capacitance at flat band conditions $C_{FB}$, the depletion depth $W_{dep}$, and the threshold voltage $V_{th}$. A distinction can be made between a quasi-static CV-measurement (QSCV-measurement) and a high frequency CV-measurement (HFCV-measurement).

In a QSCV-measurement the gate voltage is swept at a rate much lower than the recombination or generation rate of electron-hole pairs. Under these conditions the inversion layer follows the variations in the gate voltage instantaneously. In a HFCV-measurement a bias voltage is used to vary the gate voltage stepwise, and the carrier density formed in the inversion layer follows this applied bias. A small, high frequency voltage (kHz-MHz) is superposed on the bias voltage on the gate and the response to this high frequency signal is measured. The frequency is chosen sufficiently high to ensure that the free carriers are unable to follow the high frequency signal. Only the ionized impurities in the depletion area can respond to the high frequency signal. The total capacitance can be described in general by the oxide capacitance ($C_{ox}$) in series with the capacitance across the silicon also called the depletion capacitance ($C_{dep}$). The equation then becomes:
Chapter 2. Experimental system: silicon field effect transistors

\[
C = \frac{C_{ox}C_{dep}}{C_{ox} + C_{dep}}
\]  
(2.21)

The depletion capacitance depends on the depletion width. In accumulation the depletion width is zero, then \( C = C_{ox} \). Moving towards depletion the width increases and the capacitance, \( C \), decreases. When strong inversion occurs charge accumulates at the interface, while the material is still depleted. The difference between a QSCV-measurement and a HFCV-measurement is that the QSCV-measurement also measures the response of the free carriers from the inversion layer, while with a HFCV-measurement only the response of the depleted area is measured in the strong inversion region. In the strong inversion region the HFCV-measurement shows a low capacitance value of the responding dopants at the edge of the depletion area. The QSCV-measurement shows the response of the free carriers and therefore the capacitance returns to the higher oxide capacitance at the crossing to strong inversion.

In Fig. 2.5 a QSCV-measurement is shown of a MOSFET of a wafer made in June 1994 (\#4064). The HFCV-measurement of the MOSFET is not shown, because it suffers from parallel conductance, a MOS-capacitor was not available. The curve shows the typical behavior for a p-type substrate. At the negative voltage side the MOSFET is in accumulation and the capacitance is equal to the oxide capacitance \( (C_{ox}) \). The dip in capacitance is the response due to depletion, while at strong inversion the oxide capacitance is retained.

The oxide capacitance \( C_{ox} \) can be derived from the QSCV-measurement by averaging the values for the capacitance in accumulation (negative voltage side) and in strong inversion (positive voltage side). The small difference between these two values is due to a small leakage current through the oxide, which is larger in MOSFETs because of dopants in part of the oxide, then in the more simple MOS-capacitors. The oxide thickness \( d_{ox} \) is calculated using the relative dielectric constant for silicon dioxide, \( \varepsilon_r = 3.84 \) and the area of the sample, \( A = 2.70 \times 10^{-6} \text{m}^{-2} \):

\[
d_{ox} = \frac{\varepsilon_0 \varepsilon_r A}{C_{ox}}
\]  
(2.22)

The result is included in Table 2.1.

The coupling between the QSCV-measurement and the equations of Sec. 2.2.3 should be made. Following a similar approach as in the derivation of the threshold voltage, the gate voltage in the depletion and inversion
2.2. Control of carrier density

Figure 2.5: A QSCV-measurement from MOSrun June 94 with a p-type substrate.
region is related to the flatband voltage by:

\[ V_g = V_{FB} + \frac{eN_{ai}W_{dep}}{c_{ox}} + \frac{E_{BB}}{e} \]  

(2.23)

which can be rewritten with Eq. 2.9 leading to:

\[ V_g - V_{FB} = \frac{eN_{ai}W_{dep}}{c_{ox}} + \frac{eN_{ai}W_{dep}}{2c_{dep}} \]  

(2.24)

For the total capacitance one has:

\[ \frac{1}{C^2} = \frac{2(V_g - V_{FB})}{eN_{ai}\epsilon_{Si}A^2} + \frac{1}{C_{ox}^2} \]  

(2.25)

This equation expresses the relation between the capacitance and the gate voltage in the depletion and inversion region. In Fig. 2.6 a 1/C²-plot of the QSCV-measurement (Fig. 2.5) is given. The linear part in the interval of \(-0.42\) \(V\) to \(-0.30\) \(V\) reflects the linear relation of Eq. 2.25. The linear extrapolation of \(1/C^2 = 1/C_{ox}^2\) gives the flatband voltage. The flatband capacitance is determined from Fig. 2.5. The doping concentration of the substrate follows from the slope of the linear part. The Fermi level in the bulk can be calculated from Eq. 2.5. This value is needed to calculate the bandbending, the depletion width and the threshold voltage. All the numbers are listed in Table 2.1.

Using the parameters in Table 2.1 the value for the threshold voltage expected at low temperature can be calculated based on the equations of Section 2.2.4. To calculate the total bandbending the ionization energy of boron is used: \(E_{ai,B} = E_{ai,v} = 45\) meV. The value for the threshold voltage calculated for a temperature of 4.2 Kelvin is also listed in Table 2.1.

We assume that all free electrons in the depleted area are trapped in the potential well at the interface. Then the electron density at the threshold voltage is given by Eq. 2.20, and is equal to \(7.2 \times 10^{10} m^{-2}\). This number is indeed much smaller than the ionized acceptor concentration \((4.5 \times 10^{14} m^{-2})\). Therefore at \(V_g = V_{th}\) the contribution of the free electron charge can be neglected compared to the ionized acceptor concentration. Clearly the acceptor concentration is close to the effective electron concentration at the band edge, \(N_e(T)\), which might influence the calculations of the threshold voltage where the system changes from non-degenerate to degenerate. We have verified that this has a minor effect on the calculations given in Section 2.2.4.
2.2. Control of carrier density

Figure 2.6: A $1/C^2$-plot of the QSCV-measurement of Fig. 2.5
Chapter 2. Experimental system: silicon field effect transistors

Table 2.1: The experimental parameters extracted from the QSCV-measurement of MOSrun June 95. $C_{ox}$ is the measured capacitance, from which $d_{ox}$ is derived. $V_{FB}$ is extracted from the $1/C^2$-plot, and $C_{FB}$ is the corresponding capacitance of the QSCV-measurement. The doping concentration $N_a$ is calculated from the slope of the $1/C^2$-plot. The Fermi-energy $E_F(bulk)$, the maximum depletion width $W_{dep}^{max}$, and the threshold voltage $V_{th}$ are derived from the doping concentration and the flatband voltage. The expected flatband voltage is calculated from the doping concentration and the difference in workfunction. The difference between the measured and the expected value for the flatband voltage gives the density of oxide charge $q_{ox}$. Comparing the measured QSCV-curve with the expected curve calculated from $C_{ox}$ and $V_{FB}^{exp}$ gives after integration the density of interface traps $D_{it}$.

<table>
<thead>
<tr>
<th>$T(K)$</th>
<th>$C_{ox}(pF)$</th>
<th>$d_{ox}(nm)$</th>
<th>$V_{FB}(V)$</th>
<th>$C_{FB}(pF)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>626</td>
<td>147</td>
<td>-0.565</td>
<td>346</td>
</tr>
<tr>
<td>4.2</td>
<td>626</td>
<td>147</td>
<td>-0.873</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$T(K)$</th>
<th>$N_a(10^{18}m^{-3})$</th>
<th>$E_F(bulk)(eV)$</th>
<th>$W_{dep}^{max}$(um)</th>
<th>$V_{th}(V)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>1.37</td>
<td>-0.24</td>
<td>2.2</td>
<td>0.118</td>
</tr>
<tr>
<td>4.2</td>
<td>1.37</td>
<td>-0.55</td>
<td>3.3</td>
<td>0.554</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$T(K)$</th>
<th>$V_{FB}^{exp}(V)$</th>
<th>$q_{ox}(pC/m^2)$</th>
<th>$D_a(m\text{-}^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>0.33</td>
<td>$2.1 \times 10^{18}$</td>
<td>$4 \times 10^{15}$</td>
</tr>
<tr>
<td>4.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
In the ideal case the QSCV-measurement is described by two parameters: the oxide capacitance and the substrate doping concentration. Therefore the QSCV-measurement is routinely used as a check for imperfections such as fixed oxide charge, mobile oxide charge, and the density of interface traps. Shifts in the flatband voltage are an indication that oxide charge is present in the gate oxide. A change in the slope around the flatband voltage is an indication for interface traps. For the present sample the oxide charge per area is calculated by assuming that all the charge is fixed at the Si-SiO₂ interface. The value for the flatband voltage is compared with the expected flatband voltage $V_{FB}^{exp}$. The expected flatband voltage is the difference in workfunction between the gate metal and the silicon substrate. This is the value for $q_{ow}$ given in Table 2.1. It is an underestimate of the total oxide charge, in case the oxide charge is distributed through the oxide. Apart from the shift in flatband voltage the slope around the flat band voltage is changed by the traps as well, most traps are located at the interface. The total trap density can be calculated by integrating the difference between the theoretical capacitance curve calculated from the experimental value of the two parameters and the actually measured capacitance. The density of interface traps determined in this way is also listed in Table 2.1.

2.3 Requirements for optimum devices

2.3.1 Formation of a two-dimensional electron gas

For the present experiments we need samples with contacts to the inversion layer. The contacts, which are called source and drain, are made by locally n-type doping of the silicon substrate. This is done partly underneath the gate to have overlap with the region near the interface. When an inversion layer is created a channel of charge carriers is formed between the two contacts. In contrast to the formation of an accumulation layer, there is no p-n junction (Schottky barrier) formed between the channel and the contact, resulting in a low contact resistance.

The charge carriers of the inversion layer are trapped in a potential well, which is formed parallel to the Si-SiO₂ interface. The potential well is approximated by a triangular potential well. For the calculation of the energy levels the potential well $V_w$ is modeled as:

$$V_w = \begin{cases} \infty & z < 0 \\ eE_iz & z \geq 0 \end{cases}$$

(2.26)
with \( E_i \) the electric field perpendicular to the interface and \( z \) the distance from the interface into the semiconductor (see Fig. 2.7).

The electric field \( E_i \) is assumed to be constant. The electron states are then given by the Airy functions with energy levels \( E_{i;j} \):

\[
E_{i;j} = \left( \frac{\hbar^2}{2m^*} \right)^{\frac{1}{2}} \left( \frac{3}{2} \pi eE_i(j + \frac{3}{4}) \right)^{\frac{2}{3}}
\]

with \( j \) an integer denoting the different levels. However, the assumption of a constant electric field corresponds to the case when the energy levels are unoccupied. When the energy levels are occupied, then the free electrons screen the electric field. The electric field will be larger than \( E_i^{\text{min}} \):

\[
E_i^{\text{min}} = \frac{V_{th} - V_{FB}}{d_{ox}} \quad (V_g > V_{th})
\]

for the case that the gate voltage is larger than the threshold voltage. For a voltage below the threshold voltage there are hardly free carriers to screen the electric field.
2.3. Requirements for optimum devices

The spacing of the energy levels for the samples of the MOS-run June 94 mentioned in Section 2.2.5 is about 17meV. This means that at a temperature of 4.2 Kelvin (0.36meV) or below only the lowest subband will be filled, for electron densities below about $2.5 \times 10^{16} m^{-2}$. Only when the Fermi-energy exceeds this energy difference the 2nd subband will be occupied[5]. In addition we should point out that due to spin and valley splitting each energy level is 4-fold degenerate.

2.3.2 High mobility

Apart from the parameters $V_{th}$ and $c_{ox}$ of a MOSFET an important measure for the quality is the maximum value of the mobility. It provides information about to what extent charge carriers are scattered by charged impurities (Coulomb centers) or by surface roughness. It is a measure of disorder and it might fundamentally influence the crossover from a metallic state to an insulating state at low electron densities. It provides a rugged potential landscape with valleys filled with charge carriers connected by subsurface mountainpasses. The mobility is a measure of the scattering of the charge carriers by the potential landscape. The mobility ($\mu$) is defined as:

$$\mu(n_s) = \sigma(n_s) \frac{n_s e}{n_s e} \quad (2.29)$$

to relate it directly to the conductivity ($\sigma$) of the 2DEG. This relation is based on the simple Drude model for diffusive elastic scattering, which holds for metallic systems. The mobility can also be expressed in an elastic scattering time ($\tau_{el}$):

$$\mu(n_s) = \frac{e\tau_{el}(n_s)}{m^*} \quad (2.30)$$

and shows how the mobility is related to the time between scattering events from impurities and phonons.

The mobility as a function of electron density is derived from a resistance-measurement as a function of gate voltage, and given by:

$$\mu(n_s) = \frac{1}{c_{ox} \rho(n_s) [V_g(n_s) - \bar{V}_{th}]} \quad (2.31)$$

This expression can only be used when the system is in the true metallic state, which as will become clear means that it can only be used for resistance values below $\hbar/e^2$. For higher resistances the resistance is non-linear
indicating a crossover to a non-metallic system. Therefore a mobility cannot confidently be defined for low electron density.

An example of a measured curve of mobility vs. carrier density for a temperature of 1.6 and 4.2 Kelvin is shown in Fig. 2.8. The curve shows a peak in mobility of 2.61 $m^2/Vs$ for $T = 1.6$ K, and of 2.35 $m^2/Vs$ for $T = 4.2$ K. For low electron densities the mobility below the peak mobility is dominated by Coulomb-scattering of charged impurities in the sample. For higher electron density the Coulomb-scattering has less effect due to the screening of the ionized impurities by the surrounding electrons and the interface roughness scattering dominates. It is clearly visible in Fig. 2.8 that the interface roughness scattering is temperature independent (high electron density), in contrast to the Coulomb-scattering (low electron density).

The variation in peak mobility from sample to sample is within a few percent as long as the samples are from the same wafer. Also the variations

\[
\mu (m^2/Vs) \quad 1.6 K \quad 4.2 K
\]

\[
0 \quad 0.5 \quad 1 \quad 1.5 \quad 2 \quad 2.5 \quad 3 \quad 3.5 \quad 4
\]

\[
n_s (10^{16} m^-2)
\]
2.3. Requirements for optimum devices

Figure 2.9: This figure shows from top to bottom the mobility curves of the MOSruns: June 94 (#4064), May 94 (#1054), January 94 (#2014), and March 94 (#3034) at a temperature of 4.2 Kelvin.
Table 2.2: The parameters of the oven that has been changed during the cleaning procedure of the last four MOSruns.

<table>
<thead>
<tr>
<th>MOSrun</th>
<th>Jan 94</th>
<th>Mar 94</th>
<th>May 94</th>
<th>Jun 94</th>
</tr>
</thead>
<tbody>
<tr>
<td>cleaningtime</td>
<td>13 h</td>
<td>10 h</td>
<td>11 h</td>
<td>12 h</td>
</tr>
<tr>
<td>post-bake</td>
<td>1 h</td>
<td>1 h</td>
<td>2 h</td>
<td>2 h</td>
</tr>
<tr>
<td>dwelltime</td>
<td>1.55 h</td>
<td>1.30 h</td>
<td>2.15 h</td>
<td>2.15 h</td>
</tr>
</tbody>
</table>

between samples from different wafers, but processed simultaneously show hardly any deviation in peak mobility. However, the reproducibility from one run to the next is worse, indicating that we have not sufficient control over the fabrication conditions to make high mobility devices routinely. In practice we rely on repeated attempts and select the best run. Nevertheless it is a comforting thought to realize that through this procedure mobility levels are reached which are much higher than can be obtained in industrial production lines and moreover which are crucial for the present research.

In Fig. 2.9 we show the mobility for the most recent series of runs. We mention briefly a few technological factors which influence the mobility. The growth of the thermal oxide is the most critical step. Repeated oxidation is used to flatten the roughness of the wafer. The last oxide of this repeated oxidation is used as a mask for the implantation of the contacts, which is followed by an anneal to have the dopants diffused into the silicon. Before the growth of the field oxide all the oxide from the repeated oxidation step is removed. In the field oxide the pattern for the gate oxide is etched and subsequently the gate oxide is grown. The last step is expected to be the most critical step in the oxidation process, and is taken identical for all runs. The differences in recipe of the oxidation steps are minor. Except for the run of May 94, in which the chemical cleaning with ammonia and hydrogen peroxide was eliminated, there are only some changes in the duration of a few process steps. These deviations in time are not assumed to be the main reason for the variations in peak mobility. The most important is very likely the cleaning procedure of the oxidation tubes and the time between the cleaning and the actual growth, and also how frequently the tubes are cleaned. The cleaning time, the post-bake and the dwell time between cleaning and growth are listed in Table 2.2.

Unfortunately, we have not been able to identify a clearcut relationship between the variations in peak mobility from $1.5m^{-2}$ to $2.5m^{-2}$ and the
cleaning procedure. The variation in peak mobility seems to be very sensitive, but also unpredictable, except that the peak mobility is always larger than 1 $m^2/Vs$ using this recipe (Appendix A).

In order to improve the quality of the MOSFET two mutually contradictory recommendations were around and have been tried on our samples. The first was proposed by Wheeler (Yale), who had experienced in the past that the mobility can be increased by applying a negative gate voltage and heat the sample a little for several minutes. Thereafter the sample has to be cooled rapidly to low temperature with the negative gate voltage applied. The argument is to move mobile ions from the interface in the direction of the gate by applying a negative gate voltage, while the sample is heated to about 100°C. After the sample is heated for several minutes with the gate voltage applied, the mobile ions are away from the interface. Then the sample should be cooled down quickly to freeze the mobile ions at low temperature. The Coulomb scattering is expected to be less and the shift in threshold voltage due to the oxide charge is also expected to be lower.

Fig. 2.10 shows the mobility for a sample prepared by applying a negative voltage. From the figure it is clear that by applying a negative gate voltage the peak mobility decreases instead of increases. The threshold voltage is most strongly effected by this technique, it is no longer reproducible after a thermal cycle to room temperature. Also the change in threshold voltage is not reproducible from one sample to another.

The other proposal has been made by Pudalov (Chernogolovka), who suggested to apply a positive gate voltage to the sample during cool down without heating the sample in advance. The argument is that the inversion layer is then already formed before cool down, and will be more homogeneous at low temperatures. Also this method did not provide any improvement in our samples. The positive gate voltage influences the potential at the contacts, which made it impossible to take reliable measurement of the mobility for positive gate voltages. It seems that the charge of the mobile ions such as Na$^+$ accumulate in or close to the contact. The change in the potential is clearly seen when the Hall-potential is measured. It is expected to be zero in zero magnetic field, but it turned out to be of the order of the longitudinal potential, and of a random value for each contact pair. We conclude that the charge in the oxide is affected by the applied voltage, and that the reproducibility from one sample to another is poor. The suggested enhancement of the peak mobility did not appear.

None of the above methods have been used subsequently. Instead we have focused on samples as carefully made as possible.
Chapter 2. Experimental system: silicon field effect transistors

Figure 2.10: This figure shows the mobility of samples to which a negative gate voltage is applied during the cooling down of the sample to $T = 4.2K$. 
2.3. Requirements for optimum devices

2.3.3 Determination of carrier density

Although the present study focuses on electrical conduction at zero magnetic field we have used extensively data from high magnetic fields as a means to determine the carrier density more accurately. Therefore we briefly summarize a number of properties characteristic for a 2DEG in high magnetic fields where the Quantum Hall effect can be observed. The density of states of a 2DEG can be described with a stepfunction. Above an energy corresponding to the threshold voltage the density of states \( D(E) \) is:

\[
D(E) = \frac{g m^*}{2 \pi \hbar^2}
\]  

(2.32)

with \( g \) the degeneracy factor. The degeneracy factor for Si<100> MOSFETs is four, two for the spin and two for the valley splitting in the \( k_z \)-direction.

From Eq. 2.32 it follows that a linear relation between the electron density and the Fermi-energy, the equation is only valid in zero magnetic field. At high magnetic field the relation between the density of states and the Fermi-energy changes and of course also the conductance of the 2DEG. The motion of free electrons is strongly influenced by a perpendicular magnetic field. For small fields the trajectories of the electrons are only slightly bent due to the Lorentz force, but at higher fields the electrons move in circular orbits. The electrons will interfere and the quasi-classical description should be replaced by a quantum mechanical description. The Hamiltonian \( H \) is given by:

\[
H = \frac{\left(\hat{p} - e\vec{A}\right)^2}{2m^*}
\]  

(2.33)

where \( \hat{p} \) is the momentum and \( \vec{A} \) the vector potential of the magnetic field. By choosing a gauge, i.e. \( \vec{A} = B(0,x,0) \) the Hamiltonian leads to the Schrödinger equation:

\[
\left[ \frac{p_x^2}{2m^*} + \frac{p_y^2}{2m^*} + \frac{(p_y + m\omega_c x)^2}{2m^*} \right] \Psi(x,y,z) = E \Psi(x,y,z)
\]  

(2.34)

with \( \omega_c \) the cyclotron frequency \( (eBm^*) \). Eq. 2.34 is simplified by using \([p_z,H] = 0 \) and \([p_y,H] = 0 \). Then the solution for the equation can be written as:

\[
\Psi(x,y,z) = \phi(x) \exp(i k_y y + i k_z z)
\]  

(2.35)
Chapter 2. Experimental system: silicon field effect transistors

Figure 2.11: a. shows the density of states (DOS) of a 2DEG in zero magnetic field and b. in a high magnetic field with Landau levels without impurities and c. with impurities.

Now \( \phi(x) \) has to satisfy the Schrödinger equation of a harmonic oscillator. The total kinetic energy is found to be:

\[
E = (n + \frac{1}{2})\hbar\omega_c + \frac{\hbar^2k_z^2}{2m^*}
\]  

(2.36)

Evidently, the kinetic energy of the electrons in the plane perpendicular to the magnetic field is quantized. The quantization is due to quantum-mechanical interference, which affects the momentum (\( \hbar k \)) and thereby the kinetic energy. The quantized energy levels are called Landau levels. Also the density of states is changed from a constant into equidistant peaks at a distance of \( \hbar\omega_c \) (Fig. 2.11). The broadening of the Landau levels is due to the impurities in the sample.

In Fig. 2.12 the energy of the Landau levels is shown over the cross-section of a MOSFET with width \( W \), showing the rapid increase of the electrostatic potential at the boundary of the sample. The Landau levels are bent upwards and cross the Fermi-energy near the edge of the sample. In this picture, which was first used by Halperin [6] and later by Büttiker [7], there are always states available for conduction at the edge of the sample (edge channels) as long as the lowest Landau level is occupied. While conduction through the bulk only happens when the Fermi-energy is at a Landau level.

In Fig. 2.13 the longitudinal resistance of a 2DEG in a MOSFET is shown at a fixed value of the magnetic field (\( B = 3T \)). It clearly shows the oscillatory behavior called Shubnikov-de Haas oscillations due to the changing Fermi-energy through the Landau levels. Although the splitting of the Landau levels is not complete, because the resistance in the minima does not approach zero, the minima still correspond to the positions in between
Figure 2.12: The energy in the sample at the section along the Y-axis. The increase at the edge is a consequence of the changing potential at the edge, which is not included in Eq. 2.33 to Eq. 2.36.
Figure 2.13: The longitudinal resistance as a function of gate voltage (electron density) in a magnetic field.
the Landau levels. By comparing the spacing in gate voltage between the minima ($\Delta V_{g,LL}$) with the electron density per Landau level ($n_{LL}$), which is:

$$n_{LL} = \frac{ge^2B}{h}$$  \hspace{1cm} (2.37)

it is possible to derive the oxide capacitance per area, $c_{ox}$ using Eq. 2.12 leading to:

$$c_{ox} = \frac{ge^2B}{h\Delta V_{g,LL}}$$ \hspace{1cm} (2.38)

Apart from the oxide capacitance per area the minima of the oscillations contain information about the threshold voltage too. By extrapolating the series of minima to lower gate voltages one minimum corresponds to zero Landau levels filled. Measurements at different magnetic field have this minimum in common, which corresponds to the threshold voltage shown in Table 2.3.

### 2.4 Conclusions

The threshold voltage and the oxide capacitance are important parameters of the devices to determine the electron density in the inversion layer. They are determined from capacitance-voltage measurements at room temperature, as well as from Shubnikov-de Haas oscillations occurring in the longitudinal resistance at low temperatures. Both methods are in principle accurate enough. The two methods lead to different values for the threshold voltage, at $T = 4.2K$, 0.49$V$ vs. 0.554$V$ respectively. The oxide thickness of 152$nm$ vs. 147$nm$ agrees much better.

The CV-measurement is carried out at room temperature and the value at 4.2 Kelvin is derived by using the equations given in Sec. 2.2.4. The discrepancy is attributed to the rearrangement of the charges in the oxide. At room temperature part of the oxide charge is mobile, which part rearranges due to the gate voltage. It affects the flatband voltage and the threshold
Chapter 2. Experimental system: silicon field effect transistors

voltage as discussed in Section 2.2.5. At room temperature also the filling of traps affects the CV-measurements, which is less at low temperature. The Shubnikov-de Haas oscillations provide a more direct method to obtain the threshold voltage at low temperatures and is therefore used as the standard method.

The electron density calculated with Eq. 2.12:

\[ n_s = \frac{c_{ox}}{e} (V_g - V_{th}) \]  

holds also at low electron densities \((10^{14} m^{-2} - 10^{15} m^{-2})\). The electron density relevant for conduction is expected to deviate from this calculated value in the presence of localized states in the 2DEG not participating in the conduction.

The proposal to heat the device and to apply a positive gate voltage prior to cooling down did not result in improved 2DEG properties. A negative gate voltage applied to the gate without pre-heating the device leads to a decrease in mobility. More importantly, also the source-drain contacts to the 2DEG are disturbed, probably by mobile positive charge such as sodium. The best results are obtained by cooling the MOSFET without applying any voltage. Also it is found that it is best not to apply any gate voltage at room temperature before cooling the devices. All the measurements are carried out following this procedure.

References


2.4. Conclusions


Chapter 2. Experimental system: silicon field effect transistors