3 The MOSFET Transistor and the Memory Cell

Although this book addresses topics that require previous basic knowledge of the MOS transistor, an overview of the principles of such component will not be omitted. For a rigorous and complete exposition on the MOS transistor behavior, the reader should refer to specialized literature. This book will focus on certain aspects of the MOS transistor characteristics, and examples are given from the standpoint of the designer, who tends to reason in terms of potential, current, and impedance. Furthermore, the main characteristics of the Flash memory cell will be presented in read, program, and erase mode. Also in this case, the dissertation is reduced to the essential notions to understand the problems of the design. A brief history of the electrical erasing will be outlined in the case of Flash Memories, from the first generation of devices with double bias, to the current devices having a single VDD.

3.1 The MOSFET Transistor

The MOSFET transistor (or simply MOS) is defined as a voltage controlled current source. When the behavior of the device with respect to small signals is to be studied, which means small deviations with respect to a fixed steady condition (called bias or operating point), the equations of the representing model can be linearized\(^1\).

The equivalent circuit of the MOS is represented in Fig. 3.1, in which the gate terminal G is insulated, since it is supposed that no current flows into the gate, and the current generator is controlled only by means of the voltage.

The source terminal S has lower potential than the drain D in case of n-channel transistor.

---

\(^1\) There are several ways to define what is meant by “small signal”. For a MOS transistor, for example, the ratio between bias current in saturation region and signal current can be considered. The small signal condition is then \(v_p \ll (V_{GS} - V_T)\), where the voltage in lowercase refers to the signal, whereas the bias quantities are indicated in uppercase. Owing to the notion of small signal, it is possible to consider only the variations and, hence, in the analysis of the equivalent circuit, both VDD and ground are considered to be at the same potential, being stationary with respect to the signal variations.
In Fig. 3.1 $g_{ds}$ is the output conductance, $g_m$ is the transconductance since it links the variation of an input quantity, the voltage $V_{GS}$, to an output quantity, the drain current $I_{DS}$.

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}}$$  \hspace{1cm} (3.1)

$$g_{ds} = \frac{\partial I_{DS}}{\partial V_{DS}}$$  \hspace{1cm} (3.2)

The parameters described are rarely used when designing from scratch, the time when the ideas are converted into circuits. Only the simulation programs can manage them so as to solve the systems of equations and obtain the potentials of all the nodes, and the currents of all the branches. The relevant aspect is that, as it will be shown, $g_m$ is proportional to the aspect ratio (W/L), which is the only parameter that the designer can control.

What is the equivalent resistance $R_{eq}$ of a PMOS transistor connected to the supply voltage, as shown in Fig. 3.2? If we suppose that the load resistance $R$ is large enough so that the current drawn from the transistor is less than the capability of supplying current of the transistor itself, the voltage $V_{OUT}$ is very close to VDD.

$$V_{DD} \quad I_{SD} \quad V_{OUT} \quad R \quad V_{DD} \quad R_{eq} \quad I_{SD} \quad V_{OUT}$$

Fig. 3.2. Equivalent resistance for a PMOS transistor working as a current source
For example, if $V_{\text{OUT}}$ is 100 mV lower than VDD, $R_{eq}$ can be calculated by means of the Ohm’s law:

$$R_{eq} = \frac{100 \text{mV}}{I_{SD}}$$ (3.3)

For example, with $I_{DS}$ of 300 μA, $R_{eq}$ of 333 Ω is obtained.

A more accurate model that takes into account the effect of bias of the substrate (body effect) is shown in Fig. 3.3.

In this case $g_{mb}$ is defined as:

$$g_{mb} = \frac{\partial I_{DS}}{\partial V_{BS}}$$ (3.4)

![Fig. 3.3. Equivalent circuit and graphic symbol for an NMOS transistor where the substrate bias source $V_{BS}$ is highlighted](image)

The effect of the bias between bulk B and source S for a MOS transistor is quantified by means of the expression of the threshold voltage $V_T$. Conventionally, it is assumed that the conduction in a MOS transistor starts when the difference of potential between gate and source exceeds the value $V_T$ that, in case of n- and p-channel, can be calculated through the following formulae:

$$V_{T,n} = V_{T0} + \gamma \cdot \left( \sqrt{2 \cdot |\Phi_p| + |V_{SB}|} - \sqrt{2 \cdot |\Phi_p|} \right)$$ (3.5)

$$V_{T,p} = V_{T0} - \gamma \cdot \left( \sqrt{2 \cdot |\Phi_n| + |V_{SB}|} - \sqrt{2 \cdot |\Phi_n|} \right)$$ (3.6)

$$\Phi_p = \frac{kT}{q} \ln \left( \frac{N_a}{n_i} \right)$$ (3.7)

$$\Phi_n = \frac{kT}{q} \ln \left( \frac{N_d}{n_i} \right)$$ (3.8)

$$\gamma = \frac{\sqrt{2 \cdot q \cdot \varepsilon_0 \cdot \varepsilon_r \cdot N_a}}{C_{ox}}$$ (3.9)
The MOSFET Transistor and the Memory Cell

\[ q = 1.602 \times 10^{-19} \text{[C]} \]  
\( k = 8.62 \times 10^{-5} \text{[eV/K]} \)  
\( T = \text{[K]} \)  
\( e_0 = 8.854 \times 10^{-14} \text{[F/cm]} \)  
\( e_{Si} = 3.9 \)  
\( e_{SiO_2} = 3.9 \)  
\( N_a = \text{[atoms/cm}^3] \)  
\( N_d = \text{[atoms/cm}^3] \)  
\( N_i = 1.45 \times 10^{10} \text{[cm}^{-3}] \)  
\( C_{ox} = \frac{e_0 e_{Si}}{t_{ox}} \)  
\( t_{ox} = \text{[nm]} \)

\( \Phi_p \) is the p-type substrate potential, about 0.6 V, whereas \( V_{T0} \) is the threshold without the contribution of the body (i.e. with \( V_{SB} = 0 \)). For a rough estimation, it is useful to consider the variation of the threshold as proportional to the square root of the voltage between bulk and source multiplied by the body effect coefficient \( \gamma \). Such parameter ranges between 0.3 and 1.0 V\(^{1/2}\) in a typical CMOS process.

**Problem 3.1:** Identify some circuits in which the body effect is disadvantageous and others in which, on the contrary, such effect is useful. Discuss the circuits identified.

\[ V_{DS} < V_{GS} < V_{T0} \]  
\[ V_{GS} > V_{T0} \]  
\[ V_{DS} > V_{GS} - V_T \]

Fig. 3.4. Output characteristic for an NMOS transistor

In Fig. 3.4 the output characteristic of an NMOS transistor is reported. In the linear region (or triode) the following relationships exist:

\[ V_{DS} < V_{GS} - V_T \]  

(3.10)
\[ I_{DS} = \frac{1}{2} \cdot \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot \left[ 2 \cdot (V_{GS} - V_T) \cdot V_{DS} - V_{DS}^2 \right] \] (3.11)

The symbol \( \mu_n \) is the electron mobility that, in silicon, is 1417 cm\(^2\)V\(^{-1}\)s\(^{-1}\). The W/L ratio is called aspect ratio of the transistor, where W is the channel width and L is the channel length.

In the saturation region in which \( V_{DS} \geq (V_{GS} - V_T) \), Eq. (3.11) is no longer valid and the drain current can be expressed as:

\[ I_{DS} = \frac{1}{2} \cdot \mu_n \cdot C_{ox} \cdot (V_{GS} - V_T)^2 \cdot (1 + \lambda \cdot V_{DS}) \] (3.12)

The term \((1 + \lambda V_{DS})\) accounts for the fact that in saturation region the transistor characteristic is not parallel to the x-axis but, on the contrary, there is a certain slope (the value of \( \lambda \) ranges between 0.03 and 0.005 V\(^{-1}\)). The parameter \( \lambda \) is usually referred to as channel modulation parameter. Finally, the important parameters for frequency analysis are the gate, gate-to-source, and gate-to-drain parasitic capacitance.

At this point, the transconductance and output conductance of the small signal model \( g_m \) and \( g_{ds} \) can be calculated in the two working regions.

In the linear region, using Eq. (3.11) we obtain:

\[ g_m = \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot V_{DS} \] (3.13)

\[ g_{ds} = \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot (V_{GS} - V_T - V_{DS}) \] (3.14)

Starting from Eq. (3.12), the transconductance in saturation region can be calculated:

\[ g_m = \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot (V_{GS} - V_T) \cdot (1 + \lambda \cdot V_{DS}) = \frac{2 \cdot I_{DS}}{(V_{GS} - V_T)} \] (3.15)

\[ g_{ds} = \frac{\lambda \cdot I_{DS}}{(1 + \lambda \cdot V_{DS})} \] (3.16)

Similar expressions can be obtained for p-channel transistors, keeping in mind that \( V_{T,p} \) is negative and that in silicon \( \mu_p \) equals 471 cm\(^2\)V\(^{-1}\)s\(^{-1}\).

**Problem 3.2:** The name of the working regions (linear and saturation) for MOS transistors is inverted with respect to bipolar transistors: find out why.

### 3.2 Transistors Available

The transistors that the designer has at his disposal depend on the kind of process and on the performance required to realize the memory device. For a device oper-
ating at low supply voltage, for example, it is important the use of transistors having different thresholds.

Hereafter a description of the most used types of transistors will be given: p-channel transistors with thin and thick oxide, n-channel transistors fabricated in both common substrate and local triple well p-type tubs, called ip-well for sake of clarity. N-channel transistors can be LVS type (Light Voltage Shift) or NAT (Natural) depending on the kind of threshold required, and, moreover, the LVS can be either HV or LV, depending on the oxide thickness. The transistors with thin oxide have the advantage of a lower threshold voltage than those with thick oxide, but they are not able to manage voltages above VDD, and, typically, they also have shorter minimum channel length. N-channel natural transistors (NAT) have threshold voltage than is even lower than LVS, and, generally, only the low voltage version is fabricated. The HV, LV and NAT versions are also available for p-channel transistors. Table 3.1 summarizes the possible types of transistors with the corresponding symbol that will be found in the schematics of the following chapters. In the table, the values of the threshold voltage at room temperature, the oxide thickness for a typical process, the value of the body effect coefficient $\gamma$, and, finally, the value of the current that a square of transistor is able to source or sink in saturation with $V_{GS}$ and $V_{DS}$ equals to VDD are reported.\(^2\)

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol outside and inside triple well</th>
<th>$T_{th}$ typical @ 27°C [V]</th>
<th>$T_{ox}$ thickness gate oxide [Å]</th>
<th>$\gamma$ coefficient [V](^{-1})</th>
<th>$I_{DS}$ current/square [µA/square]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMOS LVS LV</td>
<td><img src="symbol1.png" alt="Symbol" /></td>
<td>0.6</td>
<td>120</td>
<td>0.95</td>
<td>360</td>
</tr>
<tr>
<td>NMOS LVS HV</td>
<td><img src="symbol2.png" alt="Symbol" /></td>
<td>0.8</td>
<td>250</td>
<td>0.96</td>
<td>300</td>
</tr>
<tr>
<td>NMOS NAT LV</td>
<td><img src="symbol3.png" alt="Symbol" /></td>
<td>-0.1</td>
<td>120</td>
<td>0.41</td>
<td>380</td>
</tr>
<tr>
<td>PMOS LVS LV</td>
<td><img src="symbol4.png" alt="Symbol" /></td>
<td>-0.6</td>
<td>120</td>
<td>0.42</td>
<td>150</td>
</tr>
<tr>
<td>PMOS LVS HV</td>
<td><img src="symbol5.png" alt="Symbol" /></td>
<td>-0.8</td>
<td>250</td>
<td>0.6</td>
<td>130</td>
</tr>
<tr>
<td>PMOS NAT LV</td>
<td><img src="symbol6.png" alt="Symbol" /></td>
<td>-1.5</td>
<td>120</td>
<td>0.42</td>
<td>250</td>
</tr>
<tr>
<td>NMOS DEP LV</td>
<td><img src="symbol7.png" alt="Symbol" /></td>
<td>-3.5</td>
<td>120</td>
<td>0.9</td>
<td>360</td>
</tr>
</tbody>
</table>

\(^2\) Such parameters are specific to each process and in this case must be considered merely as an example, to provide some figures to discuss.
The circuit design must be based on actual measurements of the transistors from recently processed wafers, to insure the correct dimensioning of the circuitry. It is important to state that the transistors do not act as ideal current sources. As we know, two well defined operating regions exist, the saturation region, in which the current supplied by the transistor $I_{DS}$ is nearly constant with respect to the voltage $V_{DS}$, and the linear (Ohmic) region, in which the behavior of the transistor is resistive.

Let’s consider a circuit like the one in Fig. 3.5 in which M1 is used to charge the capacitor $C_{LOAD}$, with $V_{IN} = VDD$; if M1 works as an ideal current source having value of $I_L$, we obtain:

$$I_{LOAD} = C_{LOAD} \cdot \frac{dV_{OUT}}{dt}$$

(3.17)

$$\Delta V_{OUT} = \frac{I_{LOAD}}{C_{LOAD}} \cdot \Delta t$$

(3.18)

with the initial condition of $V_{OUT} = 0V$ and $t = 0$. Thus, the relationship between the voltage of the output node and the charging time is linear.

![Fig. 3.5. N-channel transistor used to charge a capacitor](image)

In reality, the capacitor is initially uncharged, the current starts flowing and the output voltage increases, varying $V_{GS}$ and $V_{DS}$ of the transistor according to the load characteristic shown in Fig. 3.6. The charge finishes when M1 switches off, i.e. when $V_{OUT}$ reaches the value of VDD minus the threshold voltage of M1.

Moreover, the body effect has to be taken into account, since it causes the increase of the threshold voltage of M1 as the output voltage $V_{OUT}$ increases, slowing down the charging of the output node and stopping it at a value lower than VDD – $V_{TO}$.

**Problem 3.3:** Determine the value of $V_{OUT}$ taking into account the body effect.

Let’s analyze how the charge of $C_{LOAD}$ occurs through a p-channel transistor M2. In this case, $V_{IN} = GND$, $V_{GS}$ is fixed because it depends on VDD, not on $V_{OUT}$ like in the case of the n-channel transistor. The capacitor is charged with a constant current until $V_{DS}$ of M2 is in saturation region.
After that, in the linear region, the output voltage transient is of exponential type, considering that the transistor dynamic impedance varies for each value of $V_{DS}$ of M2. In this case, $I_{LOAD}$ characteristic is represented by the curve named $V_{GS5}$ in Fig. 3.7.

---

Fig. 3.6. M1 current during $C_{LOAD}$ charge

Fig. 3.7. Charging of a capacitor through a PMOS
Let's give another example, considering the circuit in Fig. 3.8, used as voltage reference with $V_{\text{OUT}} < V_{\text{IN}}$.

For M1, the biasing conditions are $V_{GS} = V_{\text{IN}}$ and $V_{DS} = V_{\text{OUT}}$, while for M2, $V_{GS} = (V_{\text{IN}} - V_{\text{OUT}})$ and $V_{DS} = (VDD - V_{\text{OUT}})$. M1 is in linear region, being:

$$V_{DS} = V_{OUT} < V_{IN} = V_{GS} \quad (3.19)$$

whereas M2 is in saturation

$$V_{DS} = VDD - V_{OUT}; V_{GS} = V_{IN} - V_{OUT}; VDD \geq V_{IN} \quad (3.20)$$

Recalling Eqs. (3.11) and (3.12) the currents that flow in M1 and M2 can be calculated.

$$I_{D2} = \frac{\beta_2}{2} \left( V_{IN} - V_{OUT} - V_{T,\text{nat}} \right)^2 \quad (3.21)$$

$$I_{D1} = \beta_1 \left[ (V_{IN} - V_{T,\text{nat}}) \cdot V_{OUT} - \frac{1}{2} \cdot V_{OUT}^2 \right] \quad (3.22)$$

$$\beta = \mu \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \quad (3.23)$$

Where $V_{T,\text{nat}}$ is the threshold voltage of the natural transistor considered. Equaling the two currents and assuming $V_{T,\text{nat}} = 0$, after some manipulations the following relation is obtained:

$$V_{OUT} = V_{IN} \cdot \left( 1 - \frac{1}{\sqrt{\rho + 1}} \right); \rho = \frac{\beta_2}{\beta_1} \quad (3.24)$$

Table 3.2 reports simulated and calculated values for $V_{\text{OUT}}$ at different values of $\rho$.

![Diagram](image)

**Fig. 3.8.** Voltage source dependent on $V_{\text{IN}}$ and on the size ratio between M1 and M2

**Problem 3.4:** Remake all the calculations of the theoretical $V_{\text{OUT}}$ taking into account the body effect of M2.
Table 3.2. Simulated and calculated values of $V_{\text{OUT}}$ as the size ratio of transistors M1 and M2 of Fig. 3.8 varies.

<table>
<thead>
<tr>
<th>W/L M2 [µm]</th>
<th>W/L M1 [µm]</th>
<th>$\rho$</th>
<th>$V_{\text{IN}}$ [V]</th>
<th>$V_{\text{OUT}}$ theoretical [V]</th>
<th>$V_{\text{OUT}}$ simulated [V]</th>
<th>VDD [V]</th>
</tr>
</thead>
<tbody>
<tr>
<td>20/3</td>
<td>20/3</td>
<td>1</td>
<td>5</td>
<td>1.46</td>
<td>1.3</td>
<td>5</td>
</tr>
<tr>
<td>10/3</td>
<td>20/3</td>
<td>0.5</td>
<td>5</td>
<td>0.91</td>
<td>0.8</td>
<td>5</td>
</tr>
<tr>
<td>20/3</td>
<td>10/3</td>
<td>2</td>
<td>5</td>
<td>2.1</td>
<td>2.1</td>
<td>5</td>
</tr>
<tr>
<td>3/20</td>
<td>3/20</td>
<td>1</td>
<td>5</td>
<td>1.46</td>
<td>1.1</td>
<td>5</td>
</tr>
<tr>
<td>3/40</td>
<td>3/5</td>
<td>0.125</td>
<td>4</td>
<td>0.2</td>
<td>0.1</td>
<td>5</td>
</tr>
<tr>
<td>3/40</td>
<td>3/3</td>
<td>0.075</td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>5</td>
</tr>
<tr>
<td>40/3</td>
<td>3/5</td>
<td>22</td>
<td>5</td>
<td>4</td>
<td>4.1</td>
<td>5</td>
</tr>
</tbody>
</table>

3.3 The Memory Cell

The memory cell we will deal with is the non-volatile cell, which is able to retain the information even without supply voltage. This is possible due to the insulated gate. Several versions of the non-volatile cell, based on the principle of the floating gate, exist. We will deal in detail with the cell named “T” because of the geometric shape, which constitutes the basic element of the array organization of the so-called NOR-type non-volatile memory. The picture or, better, the layout of the cell is shown in Fig. 3.9 with its basic features. It stores the single bit of information, the reading of several bits in parallel allows obtaining the bytes and the words.

The single cell structure is repeated to create the array that constitutes the bank of memory. Figure 3.10 shows an array of eight cells. As it can be noted, the contact is shared between two cells, which allows reducing the overall area by diminishing the number of contacts and, moreover, increasing the reliability of the device since the contact are critical in the fabrication process. The consequent parallelism of cells is often paid from the electrical point of view as we will see later on.

The cells that are on the same row of poly2 have also the same source junction, which is also shared with the cells of the following row. A source contact at intervals\(^3\) guarantees the connection of the source to ground or to a fixed potential.

---

\(^3\) Generally every 16 cells.
Particular attention is then paid to the side cells, the cells that constitutes the rows and columns placed at the edge of the array. As in any repetitive structure, breaking the symmetry at the end of the framework causes non-uniformity that results in memory cells that are different in terms of size and characteristics from those inside the array. Some dummy rows and columns are present but not electrically connected, with the function of reducing the edge effect for the used cells.

**Problem 3.5:** How should side rows and columns be fabricated in an EPROM cell array and in a Flash cell array?

The starting point for the analysis of a memory cell with floating gate is the equivalent capacitance one-dimensional model depicted in Fig. 3.11. In this model, merely electrostatic, the four electric terminals are coupled to the floating one by means of capacitors that can be derived assuming flat and parallel planes. Therefore, the capacitance depends on the thickness of the dielectric and on the area of overlap between floating gate and electrodes.

Considering that a certain charge $Q$ is present on the floating gate, it is possible to use the relationship among capacitance, charge, and difference of potential at the sides of a capacitor to calculate the potential of the floating gate $V_{FG}$ with reference to the external potentials:

$$Q = C_{FC} \cdot (V_{FG} - V_{CG}) + C_{S} \cdot (V_{FG} - V_{S}) + C_{D} \cdot (V_{FG} - V_{D}) + C_{B} \cdot (V_{FG} - V_{B})$$

(3.25)
Fig. 3.10. Layout of eight cells connected so as to form an NOR-type array, with the source shared between two rows of cells and the drain contact shared between two single cells.

Fig. 3.11. One-dimensional model for floating gate transistor
At this point, we can use the following definitions:

\[ C_T = C_{FC} + C_D + C_S + C_B \]  
\[ \alpha_S = \frac{C_S}{C_T}; \alpha_D = \frac{C_D}{C_T}; \alpha_B = \frac{C_B}{C_T}; \alpha_G = \frac{C_{FC}}{C_T} \]

(3.26)  
(3.27)

to obtain:

\[ V_{FG} = \alpha_S \cdot V_S + \alpha_D \cdot V_D + \alpha_B \cdot V_B + \alpha_G \cdot V_{CG} + \frac{Q}{C_T} \]

(3.28)

It is interesting to point out that the behavior of the cell is identical to the n-channel transistor having the gate terminal at \( V_{FG} \), i.e. the polarization of the electrodes. This observation is the essential principle of operation of the Flash cell.

Let’s analyze the behavior of the cell in the linear region. Substituting the voltage of the floating gate in Eq. (3.11), we get:

\[ I_D = \beta^{FG} \cdot \left( V_{FG} - V_T^{FG} \right) \cdot V_D - \frac{V_D^2}{2} \]

(3.29)

The superscript FG indicates quantities that refer to the floating gate. With the source and bulk at ground potential, it is possible to calculate the drain current of the cell with reference to the external potentials only. Substituting Eq. (3.28) into Eq. (3.29) we have:

\[ I_D = \beta^{FG} \cdot \left( \alpha_D \cdot V_D + \alpha_G \cdot V_{CG} + \frac{Q}{C_T} - V_T^{FG} \right) \cdot V_D - \frac{V_D^2}{2} \]

(3.30)

Now it is possible to define the parameters \( \beta \) and \( V_T \) with respect to the control gate, i.e.:

\[ \beta^{CG} = \alpha_G \cdot \beta^{FG} ; V_T^{CG} = \frac{1}{\alpha_G} \left( V_T^{FG} - \frac{Q}{C_T} \right) \]

(3.31)

In the case \( Q = 0 \), it is simple to give a circuit representation of the definition of the threshold voltage with respect to the control gate, by considering the capacitive divider formed by \( C_{GC} \) and \( C_T \) and imposing that the voltage of the floating gate equals the threshold voltage. It is now possible to rewrite the drain current of the cell in linear region only:

\[ I_D = \beta^{CG} \cdot \left( V_{CG} - V_T^{CG} \right) \cdot V_D + \frac{1}{\alpha_G} \left( \alpha_D - \frac{1}{2} \right) \cdot V_D^2 \]

(3.32)

**Problem 3.6:** Write the working equations of the floating gate cell in saturation region. Can the cell switch on also for gate voltage lower than \( V_T \)?
3.4. Reading Characteristics

Due to the array organization, the cells that are on the same column share the bias contact of the drain, whereas the cells placed on the same row share the same gate contact, as depicted in Fig. 3.12\(^4\).

![Figure 3.12](image)

**Fig. 3.12.** Array organization: the cells are connected in the x-direction, i.e. the row, by the gate and in the y-direction, i.e. the column, by the drain contact. The decoding univocally allows addressing each single cell of the array.

Thus, in order to program the cell shown in the figure, the gate must be biased at about 12 V and the drain at about 5 V. Obviously, in such a condition, all the cells on the same row have the gate at 12 V and all those on the same column have the drain at 5 V. Therefore, a cell having gate at ground but belonging to the same column of the cell that is being programmed has the drain at 5 V and suffers the effect of drain stress that tends to erase it. In Fig. 3.13 the impact of such effect on programmed cells for a typical process is shown.

If the columns contain many cells and all of them are to be programmed, in the worst case the maximum programming time allowed, say 250 \(\mu\)s, will be neces-

\(^4\) The figure anticipates many issues, but what is important at this point is the organization of the array that allows many cells to be connected to the same drain contact.
sary for each cell. Assuming a column of 1024 cells, the time of drain stress \( t_{ds} \) applied to the first cell of the column equals:

\[
t_{ds} = 1023 \cdot 250 \mu s \cong 0.256 s
\]  

This kind of stress should not lead to any problem but, in case the drain voltage were 5 V also during the reading phase there would be a loss of intrinsic charge due not only to the tunnel oxide retention but also to the electric field applied. The characteristic shown in Fig. 3.13 refers to a typical “good” cell, but the situation may be worse by several orders of magnitude for defective cells in terms of threshold shift.

\[
\Delta V_T \ [\text{V}]
\]

\[
V_D = 4.5\text{V} \quad V_D = 5\text{V} \quad V_D = 5.5\text{V} \quad V_D = 6\text{V}
\]

Fig. 3.13. Impact of the drain stress on a programmed cell. The deviation from the curve corresponding to zero volts indicates the percentage of charge loss. The family of curves has the drain voltage as parameter.

The effect of the drain stress can potentially occur even at low drain potentials. Thus, it is important to choose values of the drain potential that do not induce stress to the cells, which would lead to loss of information. In practice, the voltage of the drain terminal is driven to 1 V, to the maximum.\(^5\) Considering that the voltage applied between control gate and source during the reading is generally greater than 4 V, this causes the cell to operate in the linear region. It is now possible to calculate the transconductance of the cell in linear region (with constant \( V_{DS} \)) basing on Eq. (3.32):

\[
g_m = \frac{\partial I_D}{\partial V_{CG}} = \beta_{CG} \cdot V_D
\]  

It is possible to observe that the cell transconductance is independent from the content of charge of the floating gate, resulting in the same value for erased or programmed cells.

\(^5\) What is stated here will become more evident after studying the chapter that deals with the techniques to read the cells.
Let’s define $V_{T_0}$ as the threshold of the cell with $Q = 0$. The step of voltage $\Delta V_T$ with respect to the control gate, which differentiates the erased cell (logic “1”) from the programmed cell (logic “0”), can be calculated starting from Eq. (3.31):

$$\Delta V_T = V_T^{CG} - V_{T_0} = -\frac{Q}{C_{FC}}$$  \hspace{1cm} (3.35)

Therefore, the drain current of a cell in linear region results to be:

$$I_D = \beta^{CG} \cdot \left[ \left( V_{CG} - V_{T_0} - \Delta V_T \right) \cdot V_D + \frac{1}{\alpha_G} \left( \alpha_D - \frac{1}{2} \right) \cdot V_D^2 \right]$$  \hspace{1cm} (3.36)

Once the drain voltage used during reading has been fixed, the characteristics of the “1” and “0” cells result to be parallel and separated by a fixed quantity equal to $\Delta V_T$. The diagram of the Fig. 3.14 reports the typical quantities for the input characteristics on the axes. The axis of abscissas can be regarded as both $V_{GS}$ and bias voltage $VDD$, since it is suitable to drive the gate terminal to the maximum voltage available, which is generally the bias voltage.

For sake of simplicity, hereafter the superscript CG will be omitted and $V_{CG}$ will be referred to as $V_{GS}$ considering the source at ground.

**Fig. 3.14.** The characteristics of programmed or erased cells result to be parallel during the read operation

### 3.5 Programming

The writing operation, or programming for a Flash cell is carried out by means of the so-called hot electrons that are able to overcome the energy barrier corresponding to the thin oxide between the drain area and the insulated gate. Thus, the two terminals of the cell, source and drain, are used in different ways during the two operations: during the erase the electrons flow from the insulated gate to the source, whereas, during the program, the electrons flow from drain to gate. Such specialization of the two terminals allows optimizing the device for the
specialization of the two terminals allows optimizing the device for the two different operations\textsuperscript{6}. Let’s now examine the program mechanism.

Let’s consider an NMOS transistor (the Flash cell is an n-channel transistor) biased with a difference of potential between the gate and source terminals greater than its threshold voltage. In such a condition, minority carriers (electrons) are present all along the channel. If the drain junction is reverse-biased, some minority charges are drawn from the nearby region of channel and, thus, a reduction in the electron density along the channel from source to drain is determined. We define pinch-off voltage the $V_{DS}$ bias for which complete depletion of the channel is achieved at the drain junction. If the $V_{DS}$ voltage is further increased, the depleted region expands toward the source and its extremity is called pinch-off point. The presence of such a point defines the saturation condition for a transistor. In the specific case of the program operation, the cell is always in saturation region.

The longitudinal component of the electric field, always pointing from the source to the drain node, has a very intense gradient in the depleted region of the drain. Due to this unhomogeneity, the profile of the energy distribution of the electrons varies significantly along the direction parallel to the channel.

In general, conduction electrons tend to maintain thermal equilibrium with the lattice, yielding the energy that they acquire from the electric field through collisions with acoustic phonons, impurity, or other electrons. As a consequence, their average energy keeps around $(3/2) K T$, where $K$ is the Boltzmann constant and $T$ is the temperature of the lattice, while their speed varies linearly with the applied electric field. However, increasing the intensity of the field beyond a given limit, some electrons acquire more energy from the field than they can lose (no scatter mechanism is effective). In practice, these carriers are no longer in thermal equilibrium with the lattice and are called hot electrons.

In order to describe the kinetic of the hot electrons, the Fermi-Dirac energy distribution is introduced, similarly to what happens to the charge in thermal equilibrium, but with a specific associated temperature $T_e > T$. Silicon lattice, having atomic density of $5 \cdot 10^{22}$ atoms/cm$^3$, is characterized by electrons with covalent bonds (Fig. 3.15). Thermal energy allows some of these electrons to break the atomic bonds and, hence, they are free to move throughout the lattice, passing in conduction band. In a semiconductor, together with the electrons, also holes exist, i.e. absence of electrons, which can be represented as positive charge carriers, that move in the opposite way in the valence band. An important property of the electrons in the lattice is their distribution with respect to the allowed energy states, in the condition of thermal equilibrium. The Fermi-Dirac distribution function, $f_{FD}$, provides the probability that an energy status, $E$, is occupied by an electron:

$$f_{FD}(E) = \frac{1}{1 + e^{(E-E_f)/kT}} \quad (3.37)$$

where $E_f$ is the Fermi energy (or level), corresponding to a 50% probability that an electron occupies the related energy status. In an intrinsic semiconductor, i.e. not doped, the number of carriers in the conduction band (electrons) equals the num-

\textsuperscript{6} With the introduction of the array in the triple well, the erasing can take place along the entire channel length, making the specialization of the source no longer necessary.
ber of carriers in the valence band (holes), and the Fermi-Dirac function is symmetric with respect to that level. In a doped semiconductor, instead, the Fermi level shifts as a function of the type and quantity of doping present.

![Image of silicon crystal structure]

**Fig. 3.15.** Silicon crystallizes in a diamond structure, which is composed of two cubic lattices with centered sides, shifted with respect to each other by a quarter of diagonal. The different planes of the crystal and the respective Miller indexes are reported.

The idea of attributing an electronic temperature, $T_e$, greater than the lattice one so as to characterize the condition of the hot electrons is not completely correct, since temperature is a concept typical of the thermodynamic equilibrium. However, it is useful to understand the drift of the carriers in presence of strong electric fields.

Hot electrons are responsible for a series of mechanisms that, in various ways, feed the channel, substrate, and gate currents. In Fig. 3.16, the effects that contribute to the gate current are highlighted. Let’s now analyze in detail the phenomena of multiplication of the hot electrons in the channel.

Let’s considering the phenomenon of the generation of electron-hole pairs due to impact ionization. Impact ionization is a process of coulombian interaction among electrons that activates in presence of high electric field. In practice, it happens that a very energetic electron, impacting against an electron that is in the valence band, yields enough energy for it to pass in the conduction band. At the end of the interaction, both the electrons are in the conduction band while a hole has been generated in the valence band. The energy threshold for which such a
process takes place is nearly equal to $3/2$ of the silicon energy gap\(^7\). Considering that the gate current is around some nanoAmperes and the substrate current around microAmperes, it descends that, for the channel electrons, the ionization phenomenon is favored from the energy point of view with respect to their passage across the tunnel oxide. The negative charges, $e_1$, that originate due to primary multiplication, increase the population of electrons in the channel, phenomenon also referred to as C.H.E.I.A. (Channel Hot Electron Induced Avalanche). On the other hand, the ionization process provokes loss of energy. The electrons produced are re-accelerated by the longitudinal electric field and their probability of crossing the oxide voltage gap is concentrated in the final part of the channel. In the region of channel where the transverse electric field is direct toward the floating gate, there is a non-null probability that the holes produced by ionizing impacts may cross the oxide, providing a negative contribution to the gate current. It is necessary to bear in mind that they have to pass an energy barrier of 4.7 V, instead of the 3.2 V that is necessary for the electron injection and, furthermore, their effective mass is higher.

Fig. 3.16. Mechanism of charge injection onto the floating gate

As a consequence of the analyzed mechanisms, the amount of electrons (primary or generated by ionization) that are able to cross the tunnel oxide concentrates close to the drain junction. Recently, a physical mechanism known as C.H.I.S.E.L. (Channel Initiated Secondary Electron Injection) has been identified, which is able to produce an electron flow toward the gate, and whose peak of intensity is located far from the drain junction. The holes produced by the first impact ionization acquire enough energy from the electric field direct toward the substrate to start a new multiplication phenomenon. Due to the presence of a suit-

\(^7\) This derives from the momentum and energy conservation equations, assuming that all the carriers have the same final speed and the same mass.
able electric field, the electron generated by this mechanism are heated and attracted by the channel where they can be injected onto the floating gate. Obviously, in the first phase of the programming, this contribution is negligible, since it depends upon the probability that multiplication phenomena occur. However, as the programming operation goes on, the transverse electric field strongly opposes against the electron injection and, thus, the injection efficiency due to the C.H.E. and C.H.E.I.A. loses the most consistent component. The flow of the electrons produced through C.H.I.S.E.L. is reduced by a lower amount, since it is located in a portion of channel closer to the source, where the direction of the electric field is favorable. In this situation, the contribution due to the secondary electrons becomes more important in the expression of the overall gate current. The phenomenon described above can be regulated by means of the substrate bias. Increasing the difference of potential applied to the drain-substrate junction, the intensity of the transverse field below the channel is also increased, which is the fundamental element to the effectiveness of the injection mechanism of secondary electrons.

In conclusion, the ionization phenomenon in the channel causes an increase in the amount of conduction electrons and in the decrease of their average energy.

The tail of hot electrons is hence remarkably increased nearby the drain junction and, as a consequence, the charge flow toward the floating gate. During program, when the floating gate voltage reaches the imposed value, $V_{DS}$, the transverse electric field dramatically reduces the charge injection in the channel region close to the drain, causing a quick reduction in the gate current.

It is evident that the charge flow across the oxide tightly depends upon the profile of the energy distribution of the hot electrons. Despite the fact that it is very complicated to express the relationship with the voltage applied, it is evident that an increase in the drain voltage acts to increase the percentage of the electrons having energy greater than the oxide energy gap.

From the operating point of view, the minimum speed with which the device is due to accomplish the program operation is determined, and this also defines the voltage that must be applied. Supposing that the threshold voltage is increased by 3 V from an initial $V_T$ of 2.5 V, the bias configurations that fulfills the requirements can be found. The choice will select the lowest voltages, thus minimizing power consumption and electric stress.

The flow of secondary electrons across the tunnel oxide induced by the negative bias of the substrate is one of the most popular approaches to reduce the voltages applied to the cell terminals. Of course, the technological process must account for the possibility of fabricating the array in triple-well as highlighted in Fig. 3.17. By reducing the voltage of the control gate and the drain, it is possible to reduce the intensity of the electric field across the tunnel oxide, with the consequent improvement in terms of gate and drain stress (see Chap. 20).

In practice, the negative substrate bias reduces the voltage of the pinch-off point, increasing the longitudinal component of the electric field. Moreover, owing to the effect on the threshold voltage (the Flash cell is an NMOS transistor), a decrease in the substrate voltage greatly reduces the overall drain current during program, with the consequent power saving.
Fig. 3.17. The substrate of the array (ip-well) is insulated from the substrate of the rest of the circuitry by means of the n-well tub. In this way, it can be biased with a negative voltage during program to increase the writing effectiveness and diminish the drain current of the cell.

In Fig. 3.18 the characteristic of the threshold voltage step, \( \Delta V_T \), is reported versus the programming time in the case of cell with grounded substrate: this is the so-called program curve. We recall that the threshold voltage of the cells is varied by the gate current by means of charge accumulation on the floating gate.

In Fig. 3.18 two different zones can clearly be distinguished. The first in which there is a fast increase in \( V_T \) in a short time; the second in which there is a slower increase and a logarithmic function of the time. The point of separation between
the two zones corresponds to the condition in which $V_{FG}$ equals $V_{DS}$, i.e. the voltage of the floating gate equals the voltage of the drain, and, on top of that, it corresponds to the threshold voltage step also known as $\Delta V_T^\ast$.

Let’s suppose to apply a $V_{GS1}$ gate voltage and a $V_D$ drain voltage to an erased cell. The writing by means of hot electrons will continue until the voltage of the floating gate equals the drain voltage. If, on the contrary, the gate voltage applied is $V_{GS2} > V_{GS1}$, we obtain a higher initial gate voltage of the floating gate. As a consequence, in order to reach the $\Delta V_T^\ast$ condition, we have to inject more electrons onto the floating gate, which results in a higher threshold voltage of the written cell.

In the case in which $V_{FG} > V_D$, the conditions favor the charge injection onto the floating gate, whereas if $V_{FG} < V_D$, the charge that has already been injected provokes a reduction in the potential of the floating gate below the value imposed by the drain. The part of the channel closer to the drain diffusion has higher voltage than the floating gate, which does not allow the majority of the hot electrons (generated in this zone of high longitudinal electric field) to cross the oxide. Thus, the slope of the program curve in the zone where $V_{FG} < V_D$ changes, and the programming speed quickly diminishes. In terms of gate current, we obtain that $I_G$ is nearly constant when $V_{FG} > V_D$, whereas the gate current diminishes exponentially in the region where $V_{FG} < V_D$.

Substituting Eq. (3.35) in Eq. (3.28), the potential of the insulated gate can be expressed as:

$$V_{FG} = \alpha_G(V_{CG} - \Delta V_T) + \alpha_D V_D + \alpha_S V_S + \alpha_B V_B$$

(3.38)

If we want to calculate the value of $\Delta V_T^\ast$, i.e. the intrinsic voltage step, we have to insert the value of the potentials in the expression above, i.e. $V_B = V_S = 0$, and express $\Delta V_T$ in the hypothesis that $V_{FG} = V_D$

$$\Delta V_T^\ast = V_{CG} - \frac{1 - \alpha_D}{\alpha_G} V_D$$

(3.39)

The value of $\Delta V_T^\ast$ is an important indicator of the programming speed. If the minimum threshold voltage to regard the cell as programmed is greater than $\Delta V_T^\ast$, the cell will follow the knee of the program curve, going in a region where the gate current is very low, with a penalty in terms of programming speed. The problem of the maximization of $\Delta V_T^\ast$ through the variation of the physical parameters that impact on the capacitive ratios poses. What we do is increasing the overlap area between control and floating gate by means of the so-called wings. As it can

---

8 Near the beginning of the injection process, the inversion layer extends almost all the way to the drain, and the field in the oxide is attractive except for a small portion very near the drain. Current begins to flow through the oxide at the point where the electrons are the hottest and where the oxide field is most favorable. As the floating gate charges up, the floating gate-to-source voltage drops and the drain pinch-off region moves toward the source.

9 In the cells of the latest generation, such a sharp distinction between the two programming zones no longer exists.
be seen in Fig. 3.9, while the capacitive coupling between control and floating gate increases because of the wings, the coupling with the substrate is nearly constant due to the high thickness of the field oxide underneath. In this way, the $\alpha_g$ coefficient increases and, taking into account Eq. (3.39), also the value of $\Delta V_T^*$.  

When operations of modification of the charge content of the floating gate are examined, it is necessary to account for the topological distribution of the cells within the array. For example, the resistive paths that connect the source of the single cells to the ground contact present resistance ranging from a few to several thousands Ohms, depending on the size and the kind of insulation used. This resistance worsens the programming characteristics and further widens the cell distribution. Finally, temperature variations and misalignments of the masks during fabrication worsen the problem.

### 3.6 Program Algorithm

The circuitry necessary to accomplish the programming is very complex. The advent of technologies with more and more reduced dimensions imposes precise control on the potentials to apply to the cells and to timings. In the case of devices with single supply, another problem for the designer is the cell current consumption during this phase. The control circuitry that governs the operation is realized in the same way also in the case of double supply devices. Major attention is paid to the way in which the program voltages are applied. First, the gate voltage must go high in order for the channel to form, and then the drain node must be pulled up with the program pulse. If the drain went up before the gate, unpleasant accidents might occur to all the cells connected to the column, such as spurious erase or program, *snap-back* phenomena (Fig. 3.19) and degradation of the cell performances. Thus, the control circuitry verifies the correct timing application of the signal, their duration and amplitude.

The instant of application of the drain pulse, with the gate already up, causes the cell to sink very high current, a part of which charges the parasitic capacitance of the bit line, whereas the other part crosses the channel producing the necessary hot electrons. This current peak can be estimated around 1 mA in the worst case.

Single supply devices cannot afford charge pumps that, in the hypothesis that we want to program an entire word, are able to source 16 mA during program with a 5 V output voltage. The first step is the adoption of the program on a byte basis instead of on a word basis, even though this increases the programming time.$^{10}$

---

$^{10}$ The programming time directly impacts on the cost of the final system since a part of the memory is usually programmed by the manufacturer. The typical programming time ranges from 5 to 10 $\mu$s to program a word.
The MOSFET Transistor and the Memory Cell

The introduction of a control algorithm that distinguishes the two bytes that compose the word only in the case that the overall number of non-programmed bits is greater than eight may help. Even in this case, the problem of the power consumption is still present. In order to try to reduce its impact, new circuit configurations may be introduced. The first consist in programming by means of a ramp on the gate, as shown in Fig. 3.20.

![Diagram showing the snap-back effect due to the switching on of the parasitic bipolar transistor associated to each MOS. The potential applied to the drain may cause the breakdown of the junction, injecting holes into the substrate and pulling up the voltage of the base of the npn transistor indicated. When the bipolar switches on, the current does not flow along the surface, but is sunk into the ground contact of the substrate. Thus, the potential of the drain node decreases, which switches off the bipolar, and so forth. An oscillation is triggered on the drain node.](image)

**Fig. 3.19.** Snap-back effect due to the switching on of the parasitic bipolar associated to each MOS. The potential applied to the drain may cause the breakdown of the junction, injecting holes into the substrate and pulling up the voltage of the base of the npn transistor indicated. When the bipolar switches on, the current does not flow along the surface, but is sunk into the ground contact of the substrate. Thus, the potential of the drain node decreases, which switches off the bipolar, and so forth. An oscillation is triggered on the drain node.

The introduction of a control algorithm that distinguishes the two bytes that compose the word only in the case that the overall number of non-programmed bits is greater than eight may help. Even in this case, the problem of the power consumption is still present. In order to try to reduce its impact, new circuit configurations may be introduced. The first consist in programming by means of a ramp on the gate, as shown in Fig. 3.20.

![Graph showing the application of the program pulse during the gate ramp to limit the current consumption](image)

**Fig. 3.20.** Application of the program pulse during the gate ramp to limit the current consumption.
During the first phase, the gate is biased at 5 V, which is enough to create the channel and, subsequently it is pulled to the final value of about 10 V by means of a ramp. During the application of the ramp, the drain pulse is applied that, in the case of the example shown, has a minimum duration of 5 μs. In this way, the initial peak of current is reduced by acting on the value of the potential applied.

As already discussed, the placement of the array in triple-well allows biasing the cell substrate, improving the program effectiveness and, thus, reducing the value of the current sunk to about 100 μA for each cell.

3.7 Erase Operation

Let’s now examine the electrical erase operation in detail. Let’s start by considering a gated diode, which is a MOS structure where only the diffusion area beneath the gate is active (Fig. 3.21). In our case, the gated diode represents a system composed of the isolated gate and the cell source junction, where the erase phenomenon takes place.

![Gate diode structure to study the erase operation](image)

If we consider the cross section along the A-A’ dashed line and draw the energy band diagram during erase, we obtain Fig. 3.22. The source has positive potential, the gate has negative potential, whereas the body is grounded. Moreover, the cell is programmed, i.e. the floating gate is charged with electrons. Such electrons have a finite probability of crossing the energy barrier related to the gate oxide, due to the Fowler-Nordheim (FN) tunneling. The gate current associated with the FN tunneling phenomenon can be expressed as:

$$I_G = A_{FN} \cdot E_{ox}^2 \cdot \exp\left(-\frac{B_{FN}}{E_{ox}}\right)$$ (3.40)
where $E_{ox}$ is the electric field across the gate oxide (between source and floating gate), while $A_{FN}$ and $B_{FN}$ are constants.

Bands’ bending is due to the applied potentials and implies that in the valence band there are electrons having the same energy as those in the conduction band (point A in Fig. 3.23).

The possibility of such an interband electron flow depends on the donor concentration in the silicon that is responsible for the band curvature. In the examined case, the $n^+$ doping concentration of the cell source is indicated in Fig. 3.23.

![Energy band structure with the applied erase potentials](image)

**Fig. 3.22.** Energy band structure with the applied erase potentials

![Variation of the $n^+$ doping concentration along the source-to-gate direction](image)

**Fig. 3.23.** Variation of the $n^+$ doping concentration along the source-to-gate direction

The voltage drop in the silicon increases as the doping concentration decreases and, thus, an intense electric field is present in the oxide beneath the gate while, on the opposite side, at the source junction end, all the potential drop is located in the
3.7 Erase Operation

silicon. In Fig. 3.24, the region where the interband tunneling phenomenon due to the electrons that leave holes in the valence band can take place is shown. Such holes flow through the depleted region of the substrate into the ground terminal and are responsible for the current called \( I_{BBT} \) (Band-to-Band Tunnel). The gate current is around some picoAmperes, while the BBT current is three orders of magnitude larger, i.e. around some nanoAmperes per cell. Also \( I_{BBT} \) can be expressed through an expression similar to the one used for the gate current:

\[
I_{BBT} = I_s = A_{BB} \cdot E_{Si}^2 \cdot \exp \left( -\frac{B_{BB}}{E_{Si}} \right) \tag{3.41}
\]

where \( A_{BB} \) and \( B_{BB} \) are constant and \( E_{Si} \) is the electric field at the silicon surface in the source region underneath the gate. \( E_{ox} \) and \( E_{Si} \) are bound together by the continuity of the displacement vector at the surface.

\[
\varepsilon_{ox} \cdot E_{ox} = \varepsilon_{Si} \cdot E_{Si} \tag{3.42}
\]

\[
E_{ox} = 3 \cdot E_{Si} \tag{3.43}
\]

![Fig. 3.24. The region underneath the gate where the tunneling across the oxide and the interband tunneling can take place](image)

If we draw the logarithmic diagram of the BBT current as a function of the source voltage and regard the gate voltage as a parameter, we obtain Fig. 3.25. By substituting the difference of potential between the source and the gate nodes for the source voltage on the abscissa axis in Fig. 3.25, we obtain the diagram in Fig. 3.26 where the characteristics of the BBT current are overlapped.
The MOSFET Transistor and the Memory Cell

Fig. 3.25. Band-to-Band Tunnel current as a function of the source and isolated gate voltage.

\[ \ln(I_{BBT}) \]

\[ V_s \] [V]

\[ V_G \] [mA]

Fig. 3.26. BBT and gate current as a function of the difference of potential between the source node and the isolated gate.

Let’s now consider the memory cell taking into account Eq. (3.25) that associates the voltage of the isolated gate with the source and gate node potentials, the charge trapped, and the capacitive ratios. The drain does not have any impact (at least at a rough estimate) since it is left floating during the erase operation while the substrate is grounded.

\[ V_{FG} = \alpha_S \cdot V_S + \alpha_G \cdot V_{CG} + \frac{Q}{C_T} \]  \hspace{1cm} (3.44)

If we neglect the voltage drop in the floating gate and in the silicon, the electric field across the tunnel oxide can be calculated as follows:

\[ E_{ox} = \frac{|V_{FG} - V_S|}{t_{ox}} = \left[ \frac{\left(\alpha_S - 1\right) \cdot V_S + \alpha_G \cdot V_{CG} + \frac{Q}{C_T}}{t_{ox}} \right] \]  \hspace{1cm} (3.45)
3.7.1 Erasing at Constant Voltage

In this case the erase operation is carried out by leaving the drain floating, whereas the gate is grounded and the source is pulled up to 10–12 V. The necessary electric field is therefore obtained by controlling only one of the cell terminals. This mode is implemented in flash memories that, besides VDD, have a second bias voltage dedicated to program and erase operations, named VPP. At the initial time, the electric field across the thin oxide depends only upon the charge stored on the floating gate, \( Q \). The erase operation is carried out by discharging the isolated gate and, at the same time, the electric field and the current diminish in an exponential fashion with the electric field.

The problem is that the initial peak of the electric field, caused by the application of the source voltage, may provoke the breakdown of the source junction. The solution that prevents the junction damage and the cell degradation consists in applying the erase voltage to the source by means of a voltage ramp instead of applying it directly. A simple alternative is the insertion of a resistor, \( R \), to limit the voltage (Fig. 3.27).

![Circuit configuration and voltage of the source node in the case of constant erase voltage](image)

**Fig. 3.27.** Circuit configuration and voltage of the source node in the case of constant erase voltage
The load characteristic of the resistor is overlapped to the $I_{BBT}$ characteristics. With the resistor we do not erase through a ramp since the horizontal part of the characteristic, which becomes vertical later, is covered in a short time. Erase can be performed because the voltage drop on the resistor is small and allows obtaining the necessary electric field due to the fact that the gate current is much less than $I_{BBT}$.

The resistor value is determined so as to pull the source potential to a value definitely smaller than the breakdown potential at the beginning of the erase phase, when the BBT current is maximum.

As for program, we can define the erase characteristic as the curve that expresses the threshold voltage shift as a function of time. Figure 3.28 shows as the time that is necessary to reach the $V_{T,\text{end}}$ threshold voltage is independent of the $V_T$ voltage of the programmed cells at the beginning of the erase phase. Of course, this is true for a given tunnel oxide thickness.

![Figure 3.28. Threshold voltage as a function of time](image)

In order to better understand this behavior, let’s erase two cells, geometrically and electrically identical, that have different initial charge on the floating gate (Fig. 3.29), i.e. different $V_T$. During erase we suppose that the voltage applied to the terminals of the two cells is the same, which means that the two control gates are equipotential as well as the two sources since the local currents are so small that they cannot modify the voltage drop.

At the beginning, the cell having more charge and, therefore, greater threshold voltage, has a more intense electric field and is erased more quickly. The faster cell reaches the amount of charge of the slower cell and, from that moment on, they are erased together since they undergo the same electric field.
Fig. 3.29. Two identical cells with different initial amount of charge are erased simultaneously since they belong to the same sector. The slower cell waits for the faster cell so that they have the same electric field applied.

A clarification is now necessary. The tunneling phenomenon that causes the erasing is located in a very narrow region, at least in the channel direction. The polysilicon grains that make up the isolated gate have the same size of the erase region. Besides the geometrical differences, also the possible charge trapped, due to the various process steps that eventually modifies the band structure at the interface\(^1\), must be taken into account. All the variations can be regarded as an equivalent variation of the oxide thickness. It is important to notice that the shape of the threshold voltage distribution of the erased cells does not depend on the shape of the corresponding distribution of the programmed cells (before erase). It is also possible to regard such a distribution as the overlap of a gaussian distribution and a poissonian tail.

The adoption of the mixed erase instead of the source erase increases the breakdown margin that, we recall, may be a problem only if it involves few cells that undergo source junction degradation caused by the high current flow.

In single bias source devices, high voltages are internally generated by means of charge pumps (see Chap. 15). The limited availability of current sourced by such circuits has led to the negative gate erase.

Taking into account Eq. (3.45) and supposing that:

\[
\alpha_S = 0.15; \Delta V_T = 3V; T_{ox} = 10nm; E_{ox} = 10MV/cm
\]

we obtain

\[
V_{G} = -8V; V_{S} = 5V
\]

Notice that, in this case, also the gate capacitive coupling takes place. The advantage in terms of reduction of the voltage applied to the cell source is evident.

---

\(^1\) The trapped charge is responsible for the phenomenon known as “erratic bit”. During the write/erase cycles, some cells can be found whose \(V_T\) might be negative after the \(n\)-th erase cycle, and might become positive after \(m\) more cycles and eventually negative again after \(n+m\) further cycles.
3.7.2 Constant Current Erase

The main drawback of the standard erase method with constant voltages applied is the strong dependence of the erase time on $V_s$ (typically VPP or VDD) and temperature. Moreover, this approach implies a complete dependence of the electric field on the cell process variation, giving rise to a variable peak of both source and gate current, occurring at the beginning of the erase operation.

On the contrary, the constant current erase method consists in keeping the gate current constant during the whole erase operation. This is obtained by forcing a constant current in the source node in such a way to maintain the electric field always constant. It is based on the observation that, given a certain electric field, the ratio between gate and source current is constant. In fact, as clearly shown in Fig. 3.30, given a determined $I_s$ and considering Eq. (3.40) and Eq. (3.41), $E_{si}$ is fixed and also $I_g$ is known. The $I_g/I_s$ ratio is independent of the bias conditions and depends only on the electric field. Moreover, since the gate current is constant, the threshold voltage shift of the cell, proportional to the integral of $I_g$, follows a linear time law, whose slope is a function of the chosen source current. The value of the source current can be determined according to the defined erase time and maximum electrical field, chosen according to reliability considerations.

In fact, the charge trapping/generation in the oxide is a strong function of the charge flowing through it and the electric field. The higher the electric field, the greater the oxide degradation. Hence, it is possible to control the erase time degradation, which depends on the oxide damage, by controlling the maximum electric field.

Therefore, using this kind of approach, the threshold voltage variation is constant with time, differently from all the other cases in which the variation is fast at the beginning and slows down progressively. On the other hand, what happens if
\[ \Delta V \] is constant and two cells have different amount of charge? Is the distribution of the erased cells similar to the distribution of the programmed cells?

The answer is negative since the cells tend to behave uniformly, i.e. those that have less charge “wait” for those that are erased more quickly because they have more charge at the beginning. The choice of the current determines the electric field so that the erase speed is the same also in the case of different devices. The source voltage will be tuned in order for the electric field to fulfill the design value.

The main drawback is the precision with which erase is stopped. In the previous cases, the voltage step obtained though a constant pulse at the end of the erasing diminishes and, hence, the precision of the final value of the threshold voltage increases. In this case, instead, the variation of the threshold voltage is the same during all the pulses and, thus, the error with respect to the desired threshold voltage is larger.

### 3.7.3 Erasing at Negative Gate and Triple-Well Array

The fabrication of the cell array in triple-well (Fig. 3.31) allows biasing the substrate with positive voltage, eliminating the voltage drop responsible for the curvature of the energy bands that causes the spurious band-to-band tunneling current. This result has a fundamental importance for the present memories that work with a single bias voltage that is smaller than the program and erase voltages. The high voltages are generated by means of charge pumps (see Chap. 15) that typically have a limited capability to source current. Increasing the available current means increasing the size of the capacitors of the pumps, with evident repercussions on the overall device area.

![Memory array in triple-well](image)

**Fig. 3.31.** Memory array in triple-well. The array substrate (ip-well) is isolated from the rest of the circuits by an n-well tub
3.8 Erase Algorithm

After recognizing the code that the user communicates to the CUI\(^\text{12}\), i.e. the internal memory controller, to erase a sector, the preconditioning operation starts, that is all the cells are programmed. Such a preliminary operation is executed to guarantee uniform aging of the cell population so as to limit the width of the erase distribution. If we erased a sector containing cells with low \(V_T\), such cells would be over-erased and probable depleted to negative \(V_T\) while decreasing the \(V_T\) of the programmed cells. Figure 3.32 illustrates the failure mode induced by a depleted cell.

![Diagram](image)

**Fig. 3.32.** Effect of a depleted cell on the reading of a written cell belonging to the same column. The current sunk by the depleted cell simulates, on the sense amplifier point of view, the reading of an erased cell.

The row decoder selects the written cell, F1, based on the address applied by the user. This cell does not sink current even when \(V_{GS}\) equals VDD. The sense amplifier (see Chap. 12) recognizes the cell as written since the potential of node MAT equals VDD. The presence of a depleted cell, F2, with threshold voltage

\(^{12}\) It is the acronym for Command User Interface. It is a finite state machine used to decode the commands the user applies to the memory.
equal to -2 V, determines a current flow along the bit line. The potential of node MAT decreases and the reading is not correct.

Let’s go back to the erase algorithm. At this point, we have a group of programmed cells with gaussian distribution of $V_T$. During the actual electrical erase phase, the source and gate voltages of the cells are forced to the proper values with respect to the adopted erase technique (see previous section). The time counter limits the erase pulse to the typical duration of 10 ms. At the end of the pulse, the erase verify phase is carried out. Such an operation is performed through a margined read that guarantees the correct cell recognition in normal read mode.

The erase pulses continue until the correct verification of all the cells of the sector has been completed. At this point it is necessary to verify that there are no depleted cells that may induce errors during read. It is difficult to detect the depleted cells since the presence of a cell with negative $V_T$ implies that the reading of all the other cells belonging to the same columns is dominated by the current of the depleted one. Therefore, a search algorithm is applied to verify the presence of leakage current on the columns when all the rows are grounded. When a column with such an anomaly is found, the first cell is addressed and a program pulse is applied with low gate voltage, so as to slightly increase its $V_T$, without overcoming the limit used during the erase verify phase. Subsequently, the same algorithm is applied to the second cell of the same column. If no current is drawn, it means that the depleted cell was the previous one that was already recovered with the soft programming pulse; otherwise a pulse is applied to the present cell and so forth, until the end of the column. At the end of the column, the verify phase is repeated and, in the case of presence of current, the procedure is repeated with increased gate voltage, since it means that the program pulse applied to all the cells of the column was not sufficient to recover the depleted cells.

In order to simplify the algorithm controlled by the internal finite state machine, the gate voltage is used as a parameter, whereas the drain voltage is kept to the value used during the normal program phase and the same holds for the pulse duration. As we will see in Chap. 14, the gate voltage control limits the maximum threshold voltage step for a fixed time duration of the pulse.

The erase procedure has also an important feature called erase suspend that allows the user to suspend the erase operation for an indefinite time, storing the current state of the procedure, so as to be able to restart after a proper command called erase resume. Such a feature allows the external microprocessor to access a sector of the array to retrieve data, apart from the one that is undergoing erase. In this way, the penalization due to the long erase duration (around 1 s) can be partially overcome.

### Bibliography


Yohsuka Mochizucki, “Read-disturb Failure in Flash Memory at low field”, Intel reports, Nikkei Electronics Asia, pp. 35-36, (May 1993).
