Rappels sur les structures MOS et les dispositifs mémoires

I.1. Introduction

Afin de modéliser les dispositifs composés de transistors de petites dimensions ou des dispositifs plus complexes de type mémoire, il est nécessaire de rappeler le fonctionnement des dispositifs élémentaires tels que les structures MOS (Metal-oxide-semiconducteur) et de définir les paramètres qui serviront au cours de nos études. Nous commencerons ce chapitre par un rappel sur le fonctionnement et la modélisation de la capacité MOS, qui permet une approche simple des phénomènes constituant la physique du transistor MOS. Dans un deuxième temps, nous rappellerons les principales étapes de calculs, les hypothèses et les approximations qui mènent aux modèles couramment utilisés pour le transistor MOS. Enfin, une troisième partie constituera une introduction aux mémoires à grille flottante et plus particulièrement aux mémoires à piégeage discret.

Tout au long de ce document, nous considèrerons le cas de composant à substrat de type P. On peut évidemment utiliser le même formalisme pour des dispositifs à substrat de type N (en changeant les N en P et en inversant les polarités).

I.2. La capacité MOS

I.2.1. La structure

Par définition un condensateur est constitué de deux électrodes conductrices séparées par un matériau isolant. Ainsi, on appelle « capacité MIS » la superposition de trois couches de matériaux : le métal ou poly-silicium dégénéré (appelé grille), l'isolant (SiO_2, HfO_2,Ta_2O_5, Si_3N_4…), et le semiconducteur (Si, Ge…) de type N ou de type P (appelé bulk ou substrat) (cf. Fig. (I.1)).

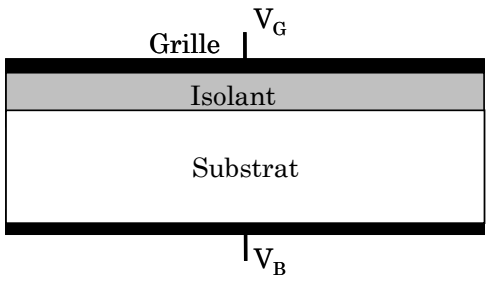

Figure I.1. *Schéma en coupe d'une structure MIS.*

La dénomination capacité MOS (pour Metal-Oxide-Semiconducteur) résulte de la nature de l'isolant qui est alors un oxyde.

I.2.2. Principe et régimes de fonctionnement

La polarisation de la capacité par une tension V_{GB}, entre la grille métallique et le substrat, implique l'apparition d'une charge Q_G dans la grille et d'une charge opposée Q_{SC} dans le semiconducteur. La variation de la tension V_{GB} modifie la valeur de ces charges, ce qui a pour conséquence les changements de régimes de fonctionnement de la capacité. La figure (I.2) présente les différents diagrammes de bandes du semiconducteur d'une capacité de type P en fonction de la tension V_{GB}.

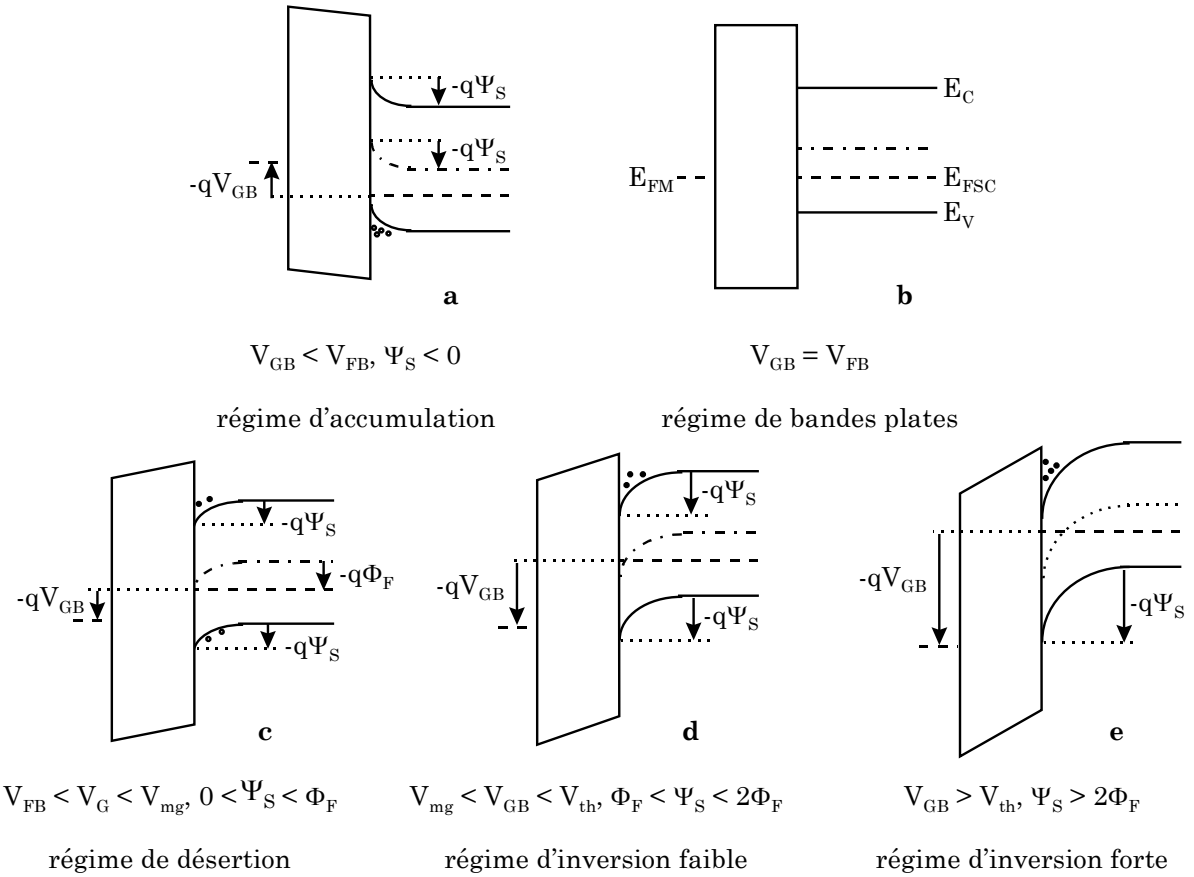

Figure I.2. *Diagrammes de bandes représentant les différents régimes du semi-conducteur en fonction du potentiel appliqué : le régime d'accumulation (***a***), le régime de bandes plates (***b***), le régime de désertion (***c***), le régime d'inversion faible (***d***) et le régime d'inversion forte (***e***).*

Ainsi une capacité MOS présente cinq régimes de fonctionnement en fonction de la tension appliquée entre sa grille et son substrat.

I.2.3. Modélisation de la capacité MOS

I.2.3.1. Les équations de bases

La capacité totale d'une capacité MOS, de surface A_{eff}, est composée de la capacité d'oxyde, C_{ox}, en série avec la capacité dynamique du semiconducteur, C_{SC} :

$$\frac{1}{C} = \frac{1}{C_{ox}} + \frac{1}{C_{SC}} \tag{I.1}$$

avec :

$$\begin{cases} C_{ox} = \dfrac{\varepsilon_{ox} A_{eff}}{t_{ox}} \\ C_{sc} = \dfrac{dQG}{d\Psi_S} = -\dfrac{dQ_{SC}}{d\Psi_S} \end{cases} \tag{I.2}$$

où Ψ_S est le potentiel de surface du substrat, et t_{ox} l'épaisseur de la couche d'oxyde. La charge au niveau de la grille, Q_G, est reliée à la tension aux bornes de l'isolant par la relation capacitive :

$$Q_G = C_{ox} V_{ox} \tag{I.3}$$

où V_{ox} est la tension appliquée aux bornes de l'oxyde.

Notons que dans les expressions (I.2), les états d'interface et la déplétion de la grille ne sont pas pris en compte.

Pour une capacité MOS, deux équations doivent être respectées : la neutralité de la charge (I.4) et la conservation de l'équation aux potentiels (I.5):

$$Q_G + Q_{ox} + Q_{SC} = 0 \tag{I.4}$$

$$V_{GB} = \Phi_{MS} + \Psi_S + V_{ox} \tag{I.5}$$

où Φ_{MS} est la différence entre les travaux de sortie de la grille et du semiconducteur et Q_{ox} la charge fixe dans l'oxyde.

A partir des équations (I.3) à (I.5), l'équation aux potentiels (I.5) s'écrit :

$$V_{GB} = V_{FB} + \Psi_S - \frac{Q_{SC}}{C_{ox}} \tag{I.6}$$

où la tension de bandes plates, V_{FB}, est définie par :

$$V_{FB} = \Phi_{MS} - \frac{Q_{ox}}{C_{ox}} \tag{I.7}$$

Notons que dans le cas d'une capacité MOS réelle, les pièges d'interface, Q_{it}, ne sont plus négligeables et la relation donnant la tension de bandes plates, V_{FB}, doit être corrigée pour prendre en compte ces charges :

$$V_{FB} = \Phi_{MS} - \frac{Q_{ox}}{C_{ox}} - \frac{Q_{it}(\Psi_S = 0)}{C_{ox}} \tag{I.8}$$

I.2.3.2. La charge du semiconducteur Q_{SC}

Exprimons, à présent, la charge du semiconducteur Q_{SC}. Celle-ci est déterminée à partir de la résolution de l'équation de Poisson, puis de l'utilisation du théorème de Gauss. Considérons N_A la concentration en atomes accepteurs ionisés (et respectivement N_D la concentration en atomes donneurs ionisés), à une dimension, pour une capacité de type P dont la concentration N_A est uniforme, l'équation de Poisson se résout simplement. Cette équation de Poisson relie la courbure des bandes du semiconducteur, $\Psi(y)$, à la densité de charges, $\rho(y)$:

$$\frac{d^2\Psi(y)}{d^2y} = -\frac{\rho(y)}{\varepsilon_{Si}} \tag{I.9}$$

où y correspond à l'axe vertical entre la surface du semiconducteur et le volume de celui-ci et $\varepsilon_{Si} = \varepsilon_{SC}\varepsilon_0$ représente la permittivité du semiconducteur.

La densité de charges dépend à la fois de la densité en porteurs libres et de la charge fixe due aux impuretés dopantes ionisées du substrat :

$$\rho(y) = q[p(y) - n(y) + N_D - N_A] \tag{I.10}$$

où p(y) et n(y) sont respectivement les densités de trous et d'électrons dans le semiconducteur.

$$\begin{cases} n(y) = n_0 \exp(\beta\Psi(y)) \\ p(y) = p_0 \exp(-\beta\Psi(y)) \end{cases} \tag{I.11}$$

où p_0 et $n_0(y)$ sont respectivement les densités de trous et d'électrons libres dans le semiconducteur loin de l'interface et β représente le potentiel thermique (q/kT),.

De plus dans le volume du semiconducteur, la condition de neutralité doit être satisfaite, c'est-à-dire $\rho\ (y \to \infty) = p_0 - n_0 + N_D - N_A = 0$ ce qui implique que $p_0 - n_0 = N_A - N_D$. L'équation (I.10) devient alors :

$$\rho(y) = -q N_A \left\{ \left(\frac{n_i}{N_A} \right)^2 \left[\exp(\beta\Psi(y)) - 1 \right] - \left[\exp(-\beta\Psi(y)) - 1 \right] \right\} \tag{I.12}$$

avec pour un substrat de type P, $p_0 = N_A$ et $n_0 = (n_i)^2 / N_A$.

A partir de l'équations (I.12) et de l'équation de Poisson (I.9), on obtient le champ électrique, $\xi(y)$:

$$\frac{d\psi(y)}{dy} = -\xi(y) = \pm\sqrt{\frac{2kTN_A}{\varepsilon_{Si}}}\left\{\left(\frac{n_i}{N_A}\right)^2\left[\exp(\beta\psi(y)) - \beta\psi(y) - 1\right] - 1 + \exp(-\beta\psi(y)) + \beta\psi(y)\right\}^{1/2} \quad \text{(I.13)}$$

En appliquant le théorème de Gauss au champ électrique à l'interface, $\iint_{(S)} \vec{\xi}\,\vec{dS} = \frac{Q_{int}}{\varepsilon_{SC}\varepsilon_0}$, la densité totale de charges dans le semiconducteur est obtenue :

$$Q_{SC} = \pm\sqrt{2kT\varepsilon_{Si}N_A}\left\{\left(\frac{n_i}{N_A}\right)^2\left[\exp(\beta\psi_S) - \beta\psi_S - 1\right] - 1 + \exp(-\beta\psi_S) + \beta\psi_S\right\}^{1/2} \quad \text{(I.14)}$$

avec un signe + si $\Psi_S < 0$ et un signe – si $\Psi_S > 0$ et N_A considéré comme constant.

I.2.3.3. La charge de la zone désertée Q_D

Pour obtenir l'expression de la charge de la zone désertée Q_D, l'équation de Poisson est résolue en omettant le terme ayant pour origine les électrons de la couche d'inversion (quantité n). La densité de charges s'écrit donc à présent :

$$\rho = q\left[p_0\exp(-\beta\Psi(y)) + n_0 - p_0\right] = q\,p_0\left[\exp(-\beta\Psi(y)) - 1 + \frac{n_0}{p_0}\right] \quad \text{(I.15)}$$

En reportant l'équation (I.15) dans l'équation de Poisson (I.9), il vient :

$$\frac{d^2\Psi}{dy^2} = -\frac{q\,p_0}{\varepsilon_{Si}}\left[\exp(-\beta\Psi(y)) - 1 + \frac{n_0}{p_0}\right] \quad \text{(I.16)}$$

En utilisant la même démarche mathématique que celle mise en œuvre pour le calcul de Q_{SC}, on obtient la charge de la zone désertée :

$$Q_D = \sqrt{2kT\varepsilon_{Si}p_0}\left[\exp(-\beta\Psi_S) + \beta\Psi_S - \frac{n_0}{p_0}\beta\Psi_S - 1\right]^{1/2} \quad \text{(I.17)}$$

Notons que puisque le substrat est de type P, la zone désertée dans le semiconducteur apparaît uniquement pour $\Psi_S > 0$, c'est pourquoi seule la racine positive de l'équation est considérée. Puisque Ψ_S est positif, il est possible de simplifier l'équation (I.17) en remarquant que :

$$\begin{cases} 1 - \frac{n_0}{p_0} = 1 - \frac{n_i^2}{N_A} \approx 1 \\ \exp(-\beta\Psi_S) << -\beta\Psi_S \end{cases} \quad \text{(I.18)}$$

La charge de la zone désertée s'exprime alors comme suit :

$$Q_D = -\sqrt{2kT\varepsilon_{Si}p_0}\left[\beta\Psi_S - 1\right]^{1/2} \quad \text{(I.19)}$$

I.2.3.4. La charge de la zone d'inversion Q_n

La charge d'inversion Q_n est définie comme la différence entre la charge du semiconducteur et la charge de la zone désertée :

$$Q_n = Q_{SC} - Q_D \quad \text{(I.20)}$$

En faible inversion, puisque $\Psi_S + V_{BS} - 2\Phi_F < 0$, alors $\exp(\beta(\Psi_S - V + V_{BS} - 2\Phi_F)) << \beta\Psi_S - 1$ du moins tant que $\Psi_S + V_{BS} << 2\Phi_F - kT/q$. Ainsi en développant Q_{SC} au premier ordre, il vient :

$$Q_{SC} \approx -\sqrt{2kT\varepsilon_{Si}p_0}\left[1 + \frac{\exp\left(\beta\left(\Psi_S - V + V_{BS} - 2\Phi_F\right)\right)}{2\left(\beta\Psi_S - 1\right)}\right]\left(\beta\Psi_S - 1\right)^{1/2} \quad \text{(I.21)}$$

A partir des relations simplifiées (I.19) et (I.21), on obtient une relation simplifiée de la charge Q_n :

$$Q_n = -\frac{1}{2}\sqrt{\frac{2kT\varepsilon_{Si}N_A}{\beta\Psi_S - 1}}\exp\left[\beta\left(\Psi_S - V + V_{BS} - 2\Phi_F\right)\right] \quad \text{(I.22)}$$

On peut également exprimer la charge d'inversion du canal en fonction du potentiel appliqué sur la grille par :

$$Q_n = Q_{SC} - Q_D = C_{ox}\left(V_{FB} - V_{GS} + V_{BS} + \Psi_S + \frac{\sqrt{2kT\varepsilon_{Si}N_A}}{C_{ox}}\left(\beta\Psi_S - 1\right)^{1/2}\right) \quad \text{(I.23)}$$

où $\beta = kT/q$ et N_A est un dopage uniforme du substrat.

I.2.3.5. La poly-désertion

Jusqu'à présent, les capacités modélisées possédaient une grille métallique (ou en poly-silicium dégénéré) ; cependant il existe des capacités dont la grille est constituée de semiconducteur non dégénéré : les capacités SOS (pour Semiconducteur-Oxide-Semiconducteur). Comme le montre la figure (I.3.a), la non dégénérescence de la grille induit une courbure des bandes d'énergie du poly-silicium près de son interface avec l'isolant. Cette courbure varie avec la polarisation de grille rendant ainsi possible l'existence des différents régimes d'un semiconducteur : accumulation, désertion, inversion faible et inversion forte. Cependant en raison des faibles niveaux de dopage de

la grille (mais encore relativement forts par rapport à ceux du substrat), il est plus probable de trouver les régimes d'accumulation et de désertion, ce dernier correspondant à ce que l'on nomme usuellement la poly-désertion (ou poly-déplétion).

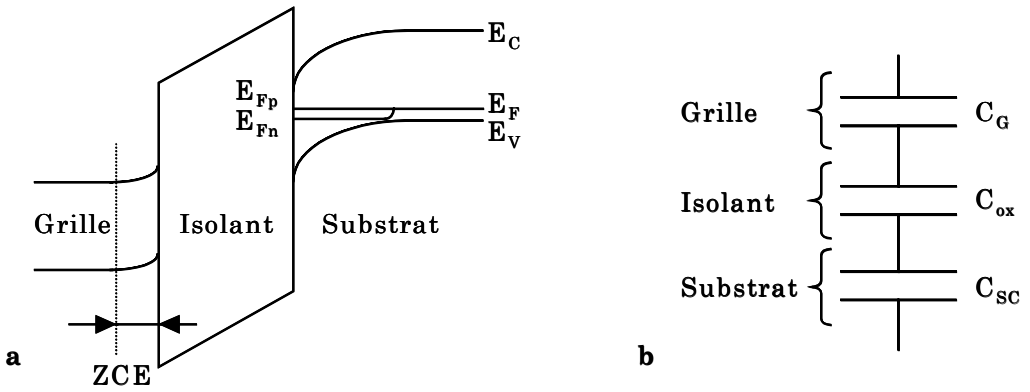

Figure I.3. *Courbures des bandes d'énergie de la structure MOS dans le cas d'une dégénérescence non complète du poly-silicium de grille (***a***) et schéma électrique capacitif équivalent (***b***).*

D'un point de vue capacitif, ce phénomène parasite s'assimile à l'apparition d'une capacité, C_G, en série avec la capacité MOS (avec un vrai métal de grille) comme l'indique la figure (I.3.b). Cela conduit à une extraction imprécise de l'épaisseur de l'oxyde de grille à partir des courbes C-V, puisque la chute de la capacité du dispositif peut être exprimée comme une augmentation de l'épaisseur de l'oxyde de grille [Huang'93].

Considérons une capacité MOS dont le substrat est de type P et la grille est en poly-silicium de type N^+. En tenant compte du potentiel de surface du poly-silicium Ψ_{SG}, l'équation aux potentiels (I.6) devient :

$$V_{GB} = V_{FB} + \Psi_S - \Psi_{SG} - \frac{Q_{SC}(\Psi_S)}{Cox} \tag{I.24}$$

Les expressions des charges en fonction des potentiels de surface sont :

$$Q_{SC} = \pm\sqrt{2kT\varepsilon_{si}p_0}\left[\frac{n_0}{p_0}\left(\exp(\beta\Psi_S)-\beta\Psi_S-1\right)-1+\exp(-\beta\Psi_S)+\beta\Psi_S\right]^{1/2} \tag{I.25}$$

$$Q_G = \pm\sqrt{2kT\varepsilon_{Si}n_{G0}}\left[\exp(\beta\Psi_{SG})-\beta\Psi_{SG}-1+\frac{p_{G0}}{n_{G0}}\left(\exp(-\beta\Psi_{SG})+\beta\Psi_{SG}-1\right)\right]^{1/2} \tag{I.26}$$

avec un signe – lorsque le potentiel de surface considéré (Ψ_S ou Ψ_{SG}) est positif et un signe + lorsqu'il est négatif. n_{G0} et p_{G0} sont les densités en porteurs majoritaires et minoritaires de la grille loin de l'interface.

I.2.3.6. Le courant tunnel Fowler-Nordheim

L'effet tunnel est un mécanisme quantique qui permet à un électron de traverser une barrière énergétique. Le mécanisme de conduction Fowler-Nordheim (FN) a été expliqué pour la première fois par Fowler et Nordheim en 1928 [Fowler-Nordheim'28] dans le cas d'émission d'électrons d'un métal dans le vide. Dans ces conditions, la barrière énergétique vue par les électrons est de forme triangulaire et les électrons peuvent la traverser par effet tunnel en se déplaçant de la bande de conduction de la cathode vers la bande de conduction de l'isolant. Cette conduction apparaît pour des structures soumises à de forts champs électriques. Il faut également préciser que les oscillations observables sur la caractéristique I-V d'une structure MIS sont dues à des effets d'interférences et de réflexions des charges aux frontières de l'isolant. La présence de charges dans l'isolant peut limiter le passage par effet tunnel dans l'isolant de la même façon qu'une zone de déplétion.

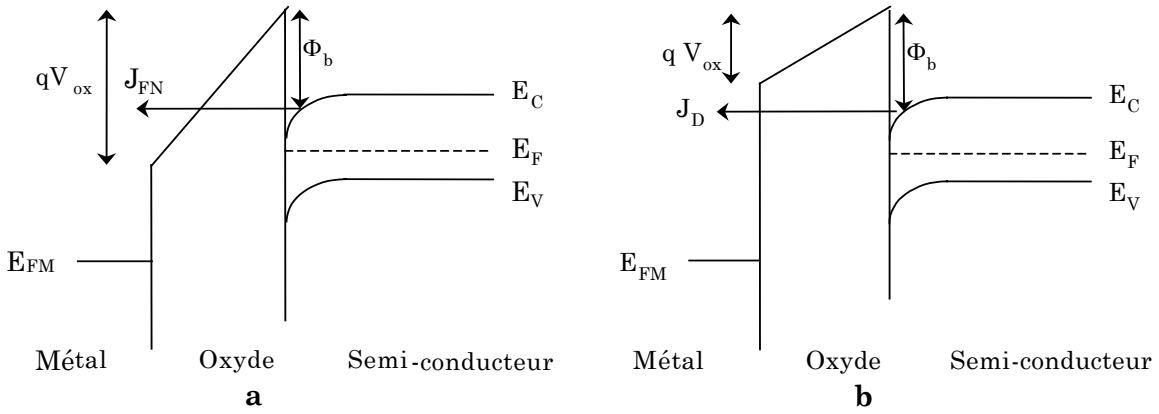

Figure I.4. *Diagramme de bandes d'une structure MOS de type P en inversion dans le cas d'un courant tunnel Fowler-Nordheim (***a***) ou d'un courant tunnel direct (***b***).*

La figure (I.4) met en évidence les deux types de transitions qui apparaissent selon la valeur de la courbure de bande de l'isolant par rapport à la hauteur de barrière, Φ_b, que les électrons voient à l'interface Si / Isolant :

- **La transition Fowler-Nordheim** pour $qV_{ox} > \Phi_b$, (figure (I.4.a)). Le champ électrique appliqué est suffisamment intense pour diminuer la largeur effective de la barrière à traverser. L'électron se retrouve alors dans la bande de conduction de l'isolant, puis il est entraîné vers l'électrode métallique.
- **La transition tunnel directe** pour $qV_{ox} < \Phi_b$, (figure (I.4.b)). Dans ce cas, le courant tunnel est dû aux électrons du semiconducteur qui traverse l'oxyde pour atteindre le métal.

La distance tunnel, d_{tun}, ainsi parcourue dépend de la hauteur de barrière Φ_b que voient les électrons à l'interface Si/Isolant et du champ électrique ξ_{ox} interne au diélectrique :

$$d_{tun} = \frac{\Phi_b}{\xi_{ox}} \tag{I.27}$$

où

$$\xi_{ox} = \frac{V_G - \Phi_{MS} - \Psi_s}{t_{ox}} \tag{I.28}$$

Dans le cas d'une structure MOS, en prenant le niveau de Fermi, E_F, comme référence des énergies le courant FN, $I_{FN}(\xi_{ox}, T)$, s'exprime de la manière suivante [O'Dweyer'73] :

$$I_{FN}(\xi_{ox}, T) = \frac{4q\pi m_M^* kT}{h^3} \int_{-\infty}^{\Phi_b} \ln\left[1 + \exp\left(\frac{-E}{kT}\right)\right] T(E)\,dE \tag{I.29}$$

où m^*_M est la masse effective de l'électron dans le métal et T (E) est le coefficient de transmission des électrons d'énergie E, à travers la barrière énergétique triangulaire.

En considérant les électrons comme un gaz à 3 dimensions, obéissant à une distribution en énergie de Maxwell-Boltzmann lorsque la longueur d'onde est négligeable devant l'épaisseur du diélectrique, alors la transparence des électrons, T(E), est calculée à partir de l'approximation de Wentzel-Kramers-Brillouin (WKB) [Fromohold'81]:

$$T(E) = \exp\left[-\frac{4\sqrt{2m_{ox}}}{3q\hbar\xi_{ox}}(\Phi_b - E)^{\frac{3}{2}}\right] \tag{I.30}$$

Les densités de courants tunnels direct et Fowler-Nordheim ont alors pour expression [Depas'95] :

$$J_{DT} = \frac{q^3}{16\pi^2\hbar\Phi_b} \frac{{\xi_{ox}}^2}{\left[1 - \left(\frac{\Phi_b - qV_{ox}}{\Phi_b}\right)^{1/2}\right]^2} \exp\left[-\frac{4}{3}\frac{(2m_{ox})^{1/2}}{q\hbar}\frac{{\Phi_b}^{3/2} - (\Phi_b - qV_{ox})^{3/2}}{\xi_{ox}}\right] \tag{I.31}$$

$$J_{FN} = \frac{q^3}{16\pi^2\hbar\Phi_b} {\xi_{ox}}^2 \exp\left[-\frac{4}{3}\frac{(2m_{ox})^{1/2}}{q\hbar}\frac{{\Phi_b}^{3/2}}{\xi_{ox}}\right] \tag{I.32}$$

où m_{ox} (exprimée en kg) est la masse effective des électrons dans l'isolant.

La figure (I.5) présente l'évolution de la densité de courant tunnel en fonction du champ électrique dans l'isolant pour les deux types de courants (direct et Fowler-Nordheim).

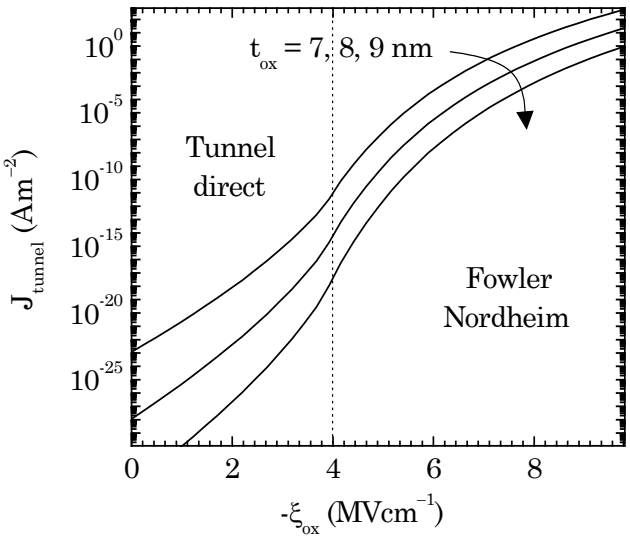

Figure I.5. *Densité de courant qui traverse une capacité MOS à substrat N^+ pour des tensions de grille positives en fonction de l'épaisseur d'isolant.*

Pour des températures proches de zéro, l'expression du courant FN d'une capacité MOS se simplifie sous la forme [Lenzlinger'69] :

$$I_{FN}(\xi_{ox}) = A_{eff}\, A\, \xi_{ox}^2\, \exp\left(-\frac{B}{\xi_{ox}}\right) \tag{I.33}$$

où A_{eff} correspond à la surface de la capacité et les coefficients Fowler-Nordheim (FN), A et B, dépendent principalement de la hauteur de barrière à l'interface oxyde/ semiconducteur et de la masse effective des électrons :

$$A = \frac{q^3}{16\pi^2\hbar\Phi_b}\frac{m_e}{m_{ox}} \tag{I.34}$$

$$B = \frac{4}{3}\frac{\sqrt{2m_{ox}}}{q\hbar}\Phi_b^{3/2} \tag{I.35}$$

où m_{ox} est la masse effective de l'électron dans l'oxyde (en général on prend $m_{ox} \approx 0.5\ m_0$).

Ces paramètres FN sont donc sensibles à la nature des électrodes et à la qualité de l'isolant notamment en terme de charges piégées. Ils se déduisent aisément à partir d'une caractéristique I-V en traçant la courbe ln (I_{FN} / A_{eff} ξ_{ox}^2) en fonction de $1/\xi_{ox}$.

I.3. Le transistor MOS à enrichissement

I.3.1. La structure

Avant de présenter les équations permettant le calcul du courant de drain du transistor MOS, il est nécessaire de définir les différentes notations utilisées [Masson'99].

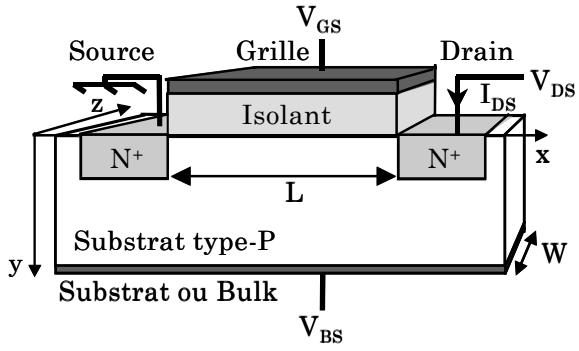

Figure I.6. *Vue schématique du transistor MOS de type N [Masson'99].*

Le transistor MOS (ou MOSFET pour transistor Métal-Oxyde-Semiconducteur à effet de champ) à canal N est un dispositif quadripolaire constitué d'une électrode de grille (G), de source (S), de drain (D) et de substrat (B) (cf. Fig. (I.6)). La longueur du transistor, notée L, correspond à la longueur de sa grille et sa largeur est notée W. La structure du transistor étant identique selon sa largeur, on le représente communément dans le plan (x,y). Nous considèrerons par la suite un transistor à canal surfacique, c'est-à-dire dont la conduction est assurée par les porteurs minoritaires du substrat (électrons dans le cas d'un NMOSFET), à l'interface entre le diélectrique de grille et le substrat.

Notons que le MOSFET possède deux électrodes supplémentaires par rapport à la capacité MOS, qui sont constituées de deux caissons dopés N^+ pour un NMOS (réservoirs à électrons). Ainsi, de nombreuses propriétés du transistor MOS découlent de celles de la capacité MOS.

I.3.2. Principe et régimes de fonctionnement

Le principe de fonctionnement du transistor MOS (ou MOSFET) repose sur la modulation d'une densité de porteurs d'une zone semi-conductrice par un champ électrique qui lui est perpendiculaire. Ce champ électrique est appliqué par l'électrode de commande (la grille) à travers un isolant (diélectrique de grille). Les porteurs créés sont des charges mobiles : électrons dans le cas d'un transistor NMOS, trous dans le cas d'un transistor PMOS. Lorsque la tension appliquée sur la grille est supérieure à une tension seuil appelée tension de seuil, notée V_T, ces charges mobiles constituent un canal de

conduction entre la source et le drain. Lorsqu'une différence de potentiel, V_{DS}, est appliquée entre la source et le drain, les porteurs affluant (côté source, de façon conventionnelle) sont collectés par le drain sous la forme d'un courant. Ainsi, de façon macroscopique, le transistor MOS se comporte comme un dispositif régulant un courant entre deux électrodes par une commande en tension.

Rappelons qu'il existe trois valeurs particulières de la tension V_{GS} :

- V_{FB} : tension V_{GS} à appliquer pour que $\Psi_S = 0$ au niveau de la source (aussi appelée tension de bandes plates).
- V_{mg} : tension V_{GS} à appliquer pour que $\Psi_S = \Phi_F$ au niveau de la source.
- V_{th} : tension V_{GS} à appliquer pour que $\Psi_S = 2\Phi_F - \Phi_C(0)$ au niveau de la source.

Notons l'apparition de l'écart entre les quasi-niveaux de Fermi, Φ_C, qui dépendent de la tension V_{DS}. En effet, les zones de drain et de source imposent un écart entre les quasi-niveaux de Fermi des électrons, E_{Fn}, et des trous, E_{Fp}, aux bornes du canal. Cet écart, Φ_C, est égal à $(E_{Fp} - E_{Fn})/q$ et prend pour valeur à la source $\Phi_C(0) = V_{SB}$ et au drain $\Phi_C(L) = V_{DB} - V_{SB}$. Le substrat étant de type P, le quasi-niveau de Fermi des trous, E_{Fp} est égal au niveau de Fermi dans le volume du semiconducteur, E_F, et ne varie pas le long du canal : seul le niveau énergétique E_{Fn} varie (cf. Fig. (I.7)).

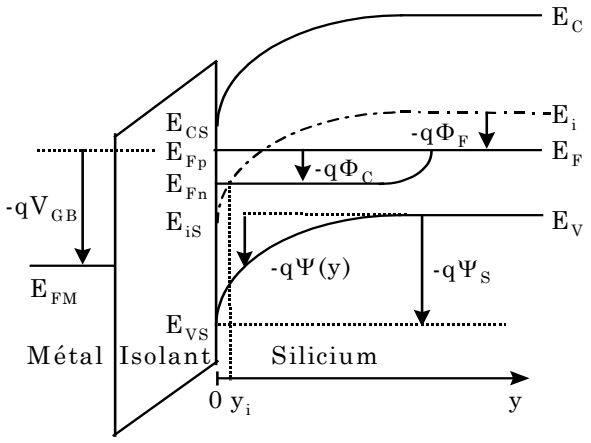

Figure I.7. *Diagramme de bandes du transistor MOS en régime d'inversion forte suivant l'axe y en un point quelconque du canal [Masson'99].*

La courbure des bandes d'énergie du semiconducteur est notée $\Psi(y)$ et la courbure totale correspond au potentiel de surface, Ψ_S. Le choix du sens des flèches a pour origine la tension que l'on applique entre la grille et le substrat. Cela revient à faire la différence entre les niveaux de Fermi du métal et du semiconducteur.

Le potentiel de volume du semiconducteur Φ_F a pour expression [Sze'81] :

$$\Phi_F = \frac{kT}{q}\ln\left(\frac{N_A}{n_i}\right) = \frac{1}{\beta}\ln\left(\frac{N_A}{n_i}\right) = -\frac{1}{q}\left(E_F - E_i\right) \qquad (I.36)$$

I.3.3. Modélisation du transistor MOS

La connaissance des équations de modélisation de la conduction dans le transistor MOS est nécessaire pour l'extraction des paramètres de fonctionnement comme la tension de seuil V_T, la mobilité à faible champ μ_0 ou la transconductance du canal g_m. Parmi les modèles décrivant les propriétés de conduction d'un transistor MOS, les modèles de Pao et Sah [Pao'66] et en feuillet [Brews'78], basés sur le principe de dérive-diffusion, permettent la continuité du courant I_{DS} entre les différents régimes de fonctionnement du transistor MOS (c.a.d. les régimes d'inversion faible, d'inversion forte, ohmique, quadratique et saturé). Ainsi, nous avons choisi d'utiliser ces deux modèles, qui reposent sur le calcul du potentiel de surface (le long du canal ou à ses extrémités).

I.3.3.1. Le modèle de Pao et Sah [Pao'66]

Le modèle de Pao et Sah [Pao'66] décrit le courant de drain en distinguant ou non les termes de conduction et de diffusion :

$$I_{DS} = -\frac{W}{L}\mu_0 \int_{\Psi_S(0)}^{\Psi_S(L)} Q_n d\Psi + \frac{W}{L}\mu_0 \frac{kT}{q}\left[Q_n(L) - Q_n(0)\right] = -\frac{W}{L}\mu_0 \int_{\Phi_C(0)}^{\Phi_C(L)} Q_n d\Phi_C \qquad (I.37)$$

où Q_n représente la charge de la zone d'inversion (par unité de surface).

D'un point de vue pratique, le calcul du courant de conduction nécessite la connaissance, à V_{GB} donnée, de la variation de la charge d'inversion et du potentiel de surface le long du canal. La relation aux potentiels liant les potentiels aux charges s'écrit :

$$V_{GB} = V_{FB} + \Psi_S - \frac{Q_{SC}(\Psi_S, \Phi_C)}{C_{ox}} - \frac{Q_{it}}{C_{ox}} \qquad (I.38)$$

Afin de déterminer Ψ_S le long du canal, à l'aide de l'équation (I.38), on considère par exemple une vingtaine de valeurs de Φ_C entre la source et le drain (c.a.d. [– V_{BS}, $V_{DS} - V_{BS}$]). La charge Q_n est alors calculée pour chaque Ψ_S, puis l'intégrale de l'équation (I.37) est évaluée par la méthode des rectangles ou des trapèzes. Notons que la détermination du potentiel de surface en un point quelconque du canal nécessite la connaissance des charges dans la structure MOS. Tandis que la charge Q_D (charge hors électrons) reste identique à celle d'une capacité MOS (I.17), l'équation de la charge du semiconducteur Q_{SC} (I.14), doit être légèrement modifiée pour tenir compte des quasi-niveaux de Fermi :

$$Q_{SC} = \pm\sqrt{2kT\varepsilon_{Si}N_A}\left\{\left(\frac{n_i}{N_A}\right)^2\left[\exp(\beta\Psi_S - \beta\Phi_c) - \beta\Psi_S - \exp(-\beta\Phi_c)\right] - 1 + \exp(\beta\Psi_S) + \beta\Psi_S\right\}^{1/2} \quad \text{(I.39)}$$

avec un signe + si $\Psi_S < 0$ et un signe – si $\Psi_S > 0$ et N_A considéré comme constant.

Quelle que soit la valeur de la tension de drain, les simulations du courant I_{DS} présentées aux figures (I.8.a) et (I.8.b) montrent qu'en régime d'inversion faible, le courant I_{DS} résulte d'un phénomène de diffusion de porteurs dans le canal, tandis qu'en régime d'inversion forte le courant de drain est presque égal au courant de conduction. De plus, la représentation en échelle semi-logarithmique de la courbe I_{DS} en fonction de V_{GB}, est linéaire en régime d'inversion faible. Cette portion de droite porte le nom de pente sous le seuil.

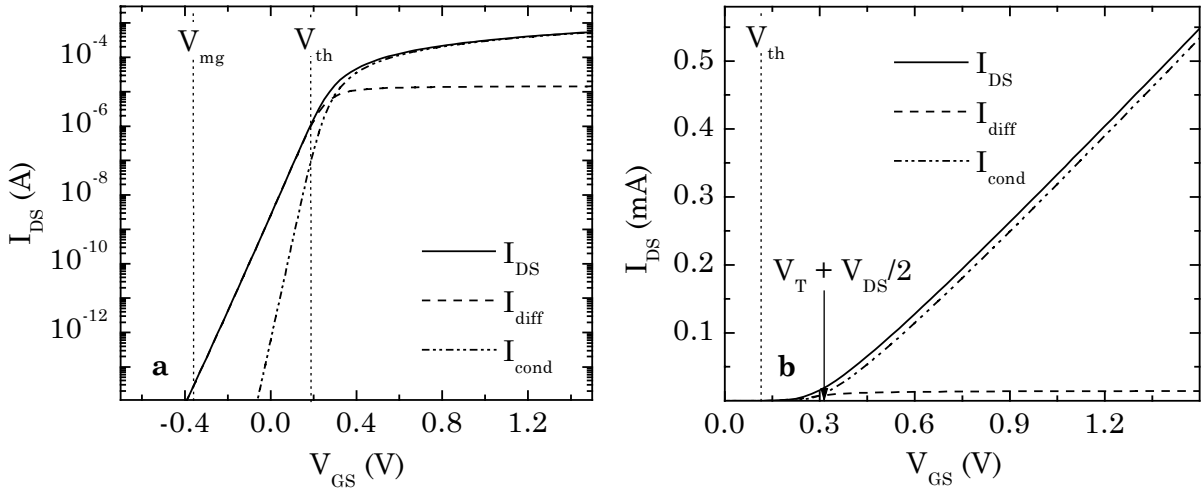

Figure I.8. *Evolution des courants de conduction et de diffusion ainsi que du courant total en fonction du potentiel de grille en échelle semi-logarithmique (***a***) ou linéaire (***b***). Les paramètres de la simulation sont : V_{DS} = 0.05 V, V_{BS} = 0 V, $N_A = 7\times10^{23}$ m^{-3}, $\mu_0 = 300$ $Vs^{-1}cm^{-2}$, L = 0.5 µm, W = 1 µm, V_{FB} = –1 V, V_{mg} = – 0.36 V et V_{th} = 0.18 V [Masson'99].*

De plus, comme le montre la figure (I.8.b), la notion de tension de seuil du transistor MOS, V_T, est différente de celle notée V_{th}. Usuellement, on considère que la tension V_T correspond au déblocage du transistor et donc à la création de la charge d'inversion (ce qui est une approximation). Elle se situe à l'intersection de la partie quasi-linéaire de la courbe avec l'axe V_{GS}. Cette notion de tension de seuil est indispensable pour des mémoires non volatiles puisque c'est elle qui représente l'information stockée. Ainsi, le calcul du courant nécessite le découpage du canal en petits éléments dont on connaît la charge $Q_n(x)$ mais pas la localisation x puisque le découpage a été fait selon Φ_C le long du canal. En supposant que, pour une polarisation donnée, le courant de drain est à flux

conservatif (c.a.d. que le courant est identique en tout point du canal), il est possible de déterminer la localisation (en x) de la charge Q_n et par suite celle du potentiel de surface $\Psi_S(x)$ et de l'écart entre les quasi-niveaux de Fermi $\Phi_C(x)$. Soit x, la distance à partir de la source, le courant de drain peut s'écrire :

$$I_{DS} = -\frac{W}{x}\mu_0 \int_{\Phi_C(0)}^{\Phi_C(x)} Q_n d\Phi_C \tag{I.40}$$

En divisant l'équation (I.37) par l'équation (I.40), on aboutit à l'équation (I.41) :

$$x = L \frac{\int_{\Phi_C(0)}^{\Phi_C(x)} Q_n d\Phi_C}{\int_{\Phi_C(0)}^{\Phi_C(L)} Q_n d\Phi_C} \tag{I.41}$$

Cette approche de type Pao et Sah présente l'avantage d'obtenir une localisation des différentes grandeurs physiques le long du canal (Q_n, Φ_C, Ψ_S). Elle autorise aussi la prise en compte d'un grand nombre d'effets parasites tels que : la présence de pièges dans l'isolant ou à son interface, la poly-déplétion de la grille, etc… La charge Q_n peut aussi être obtenue pour des cas particuliers : comme pour les effets quantiques [Masson'02] ou un dopage (vertical) non uniforme du substrat, comme nous l'expliquerons au chapitre II. Cependant, en raison du découpage en quasi-niveaux de Fermi, le calcul du courant est relativement long et ne prend pas en compte les effets 2D le long du canal. Enfin, la précision du calcul dépend du découpage du canal en quasi-niveaux de Fermi le long du canal notamment en régime de saturation.

I.3.3.2. Le modèle en feuillet [Brews'78]

En 1978, Brews donne également une expression du courant valable de l'inversion faible à l'inversion forte avant saturation en décrivant le courant de drain I_{DS} comme la somme de deux contributions : le courant de conduction et celui de diffusion [Brews'78]. Ce modèle ne nécessite pas la détermination de la charge d'inversion le long du canal puisque le calcul se fait aux frontières du canal (c.a.d. le drain et la source). La résolution de l'équation (I.37) nécessite la connaissance de la primitive de Q_n par rapport à Ψ_S qui peut être calculée à partir de l'expression de Q_n donnée par l'équation (I.23). Ainsi les équations (I.37), et (I.23) nous amènent à écrire l'équation du courant sous la forme suivante :

$$I_{DS}=\frac{W}{L}\mu_0 C_{ox}\left[\left(V_{GS}-V_{FB}-V_{BS}\right)\left(\Psi_S-\frac{1}{\beta}\right)-\frac{\Psi_S}{\beta}-\frac{\gamma}{\beta}\left(\beta\Psi_S-1\right)^{1/2}-\frac{1}{2}\Psi_S{}^2-\frac{2}{3}\gamma\left(\beta\Psi_S-1\right)^{3/2}\right]_{\Psi_S(0)}^{\Psi_S(L)} \quad (I.42)$$

où

$$\gamma=\frac{\sqrt{2kT\varepsilon_{Si}p_0}}{C_{ox}} \quad (I.43)$$

L'intégration de la charge d'inversion le long du canal aboutit à :

$$I_{DS}=\frac{W}{L}\mu_0 C_{ox}\left[F(L)-F(0)\right] \quad (I.44)$$

où la fonction F est donnée par :

$$F(x)=\left(V_{GS}-V_{FB}-V_{BS}\right)\left(\Psi_S(x)-\frac{1}{\beta}\right)+\frac{\Psi_S}{\beta}+\frac{\gamma}{\beta}\left(\beta\Psi_S(x)-1\right)^{1/2}-\frac{1}{2}\Psi_S(x)^2-\frac{2}{3}\gamma\left(\beta\Psi_S(x)-1\right)^{3/2} \quad (I.45)$$

Ce modèle, décrit plus précisément dans le livre de Tsividis [Tsividis'99], nécessite la connaissance des potentiels de surface uniquement aux bornes du canal contrairement à l'approche de Pao et Sah. Ainsi le calcul est beaucoup plus rapide. Cependant, ce gain en temps se fait au détriment de la possibilité de prendre en compte la plupart des effets parasites. De plus, ce modèle est pseudo-2D, donc ne prend pas en compte les effets 2D. Comme dans l'approche de Pao et Sah, il est possible de déterminer l'évolution du potentiel de surface et l'écart entre les quasi-niveaux de Fermi le long du canal en considérant que le courant de drain est à flux conservatif avec l'équation suivante :

$$X=L\frac{F(x)-F(0)}{F(L)-F(0)} \quad (I.46)$$

Les modélisations des structures MOS que nous venons de rappeler sont nécessaires pour la modélisation des dispositifs mémoires, basés sur le potentiel des technologies MOS. Pour l'application de nos travaux, nous nous sommes intéressés aux mémoires non volatiles à stockage discret brièvement décrite au paragraphe (§ I.4.3).

I.4. Les mémoires non volatiles

I.4.1. Généralités

Par définition, la mémoire est la propriété de conserver et de restituer des informations. Cependant, en microélectronique, il existe deux moyens pour obtenir cette

propriété. La figure (I.9) donne une classification des principales mémoires MOS qui sont traditionnellement classées en deux grandes familles :

- Les mémoires vives, désignées par le sigle générique RAM (pour Random Access Memory), c'est-à-dire mémoires à accès aléatoire ; ce sont des mémoires dans lesquelles on peut, à tout moment, écrire ou lire des informations, et ce, tant que l'alimentation électrique est présente.
- les mémoires mortes ou ROM (Read-Only Memory) sont des mémoires qui ne peuvent être que lues à partir du moment où les informations y ont été écrites; en revanche, elles possèdent la propriété de garder l'information très longtemps (spécification typique : 10 ans), même en l'absence d'alimentation électrique.

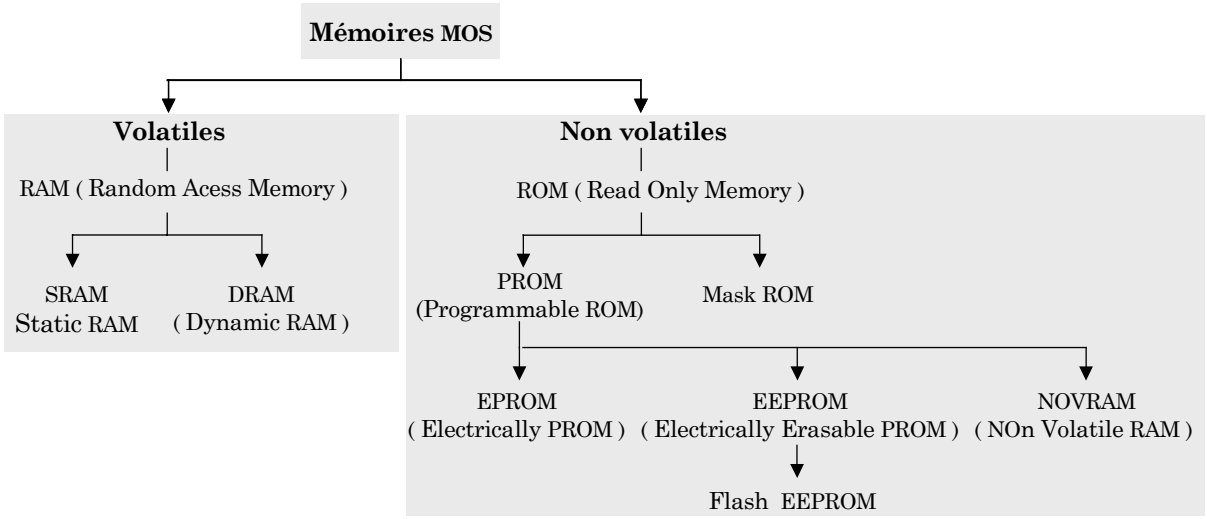

Figure I.9. *Classification des principales mémoires MOS.*

En 1967, D. Kahng et S.M. Sze [Kahng'67] présentaient la première mémoire MOS non volatile, composée d'un transistor MOS dont la grille était remplacée par un empilement de couches conductrices et non conductrices. De nos jours, les mémoires non volatiles sont quotidiennement présentent dans notre vie avec les cartes bancaires, les téléphones mobiles, les décodeurs de télévision, les ordinateurs personnels, la gestion des moteurs automobiles et beaucoup d'autres applications nécessitant la sauvegarde de l'information de façon permanente même après rupture de l'alimentation. Dans la suite du manuscrit, nous nous intéresserons uniquement aux mémoires non volatiles.

Les mémoires ROM (Read Only Memory) sont destinées uniquement à être lues, et sont essentiellement utilisées pour les jeux vidéo. Elles sont programmées, soit lors de la fabrication (activation ou non d'un transistor par masquage), soit par l'utilisateur avec des structures à base de fusibles. Le fonctionnement d'une ROM est basé sur celui du point mémoire qui est généralement constitué d'un transistor NMOS (ayant une grille

flottante) adressé en lecture par une ligne de bit connectée au drain et une ligne de mot (world line) connectée à la grille (cf. Fig. (I.10)).

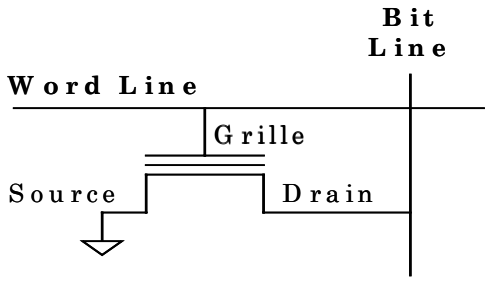

Figure I.10. *Schématisation d'un point mémoire.*

L'EPROM (Erasable Programmable Read Only Memory) permet d'effacer et de programmer les cellules. Elles sont largement utilisées en bureautique (dans les photocopieurs, les imprimantes lasers, où elles contiennent les différentes polices de caractères, dans les automates programmables, etc …). L'écriture se fait par stockage d'électrons dans une grille isolée. L'opération d'effacement par rayons Ultra Violets (UV) des EPROMs reste néanmoins lourde à mettre en œuvre : elle suppose un démontage du boîtier de son support et un passage de 15 à 20 minutes sous rayons UV. De plus, les EPROMs utilisent des boîtiers coûteux à fenêtre de quartz pour permettre ce type d'effacement. Le principal problème de fiabilité de ce type de mémoire est la rétention de l'information stockée, car le nombre de cycles d'écriture / effacement reste faible.

Les EEPROMs (Electrically Erasable PROM), développées dans le milieu des années 1970, répondent au problème de l'effacement UV par un effacement bit par bit de type électrique qui évite de retirer le circuit du système électronique pour reprogrammer la mémoire. Les EEPROMs sont en partie dédiées aux applications militaires ou spatiales. Comme le montre la figure (I.11), les EEPROMs utilisent une surface équivalente à deux transistors par cellule mémoire [Yaron'82] : le premier est utilisé comme transistor de sélection et le second est l'élément de stockage. La cellule mémoire EEPROM est traditionnellement réalisée en technologie FLOTOX (« FLOting gate Thin OXide »). Le point critique est l'utilisation d'un oxyde de grille très mince qui sépare le drain de la grille flottante (faible rendement de fabrication) et une surface occupée importante. Sous l'effet d'un champ électrique intense de l'ordre de 10 $MVcm^{-1}$, des électrons passent par effet tunnel à travers cet oxyde mince, du drain vers la grille flottante ou inversement suivant le sens du champ électrique. Ainsi cette injection d'électrons fait varier la quantité de charges de la grille flottante ce qui modifie la tension de seuil du transistor.

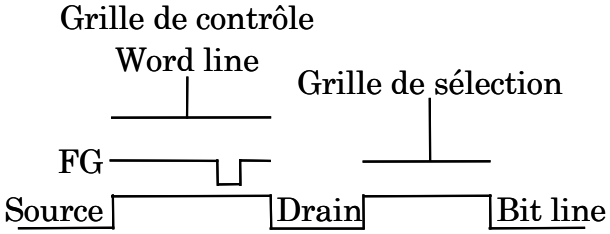

Figure I.11. *Schéma équivalent de la cellule EEPROM composée du transistor d'état en série avec le transistor de sélection.*

L'apparition des mémoires Flash EEPROMs est issue de la course aux réductions de dimensions. L' utilisation d'un seul transistor par cellule mémoire a permis un gain de place et de rapidité avec la possibilité de re-programmer les mémoires PROM et par conséquent un gain en terme de coût de production. Le terme Flash traduit le fait que les données d'un bloc entier sont effacées d'un seul coup. Actuellement, les mémoires Flash représentent la famille la plus importante des mémoires non-volatiles en raison de leur grande densité d'intégration, de leur rapidité d'écriture et de lecture [Pavan'97].

I.4.2. Les mémoires Flash

La première mémoire flash fut présentée en 1984 par Masuoka [Masuoka'84]. Décrivons la structure et le fonctionnement de ces mémoires Flash.

I.4.2.1. La structure des mémoires Flash

La mémoire Flash est constituée d'un transistor MOS dont la structure de la grille a légèrement été modifiée avec une grille de contrôle et une grille flottante (FG) emprisonnée dans l'isolant.

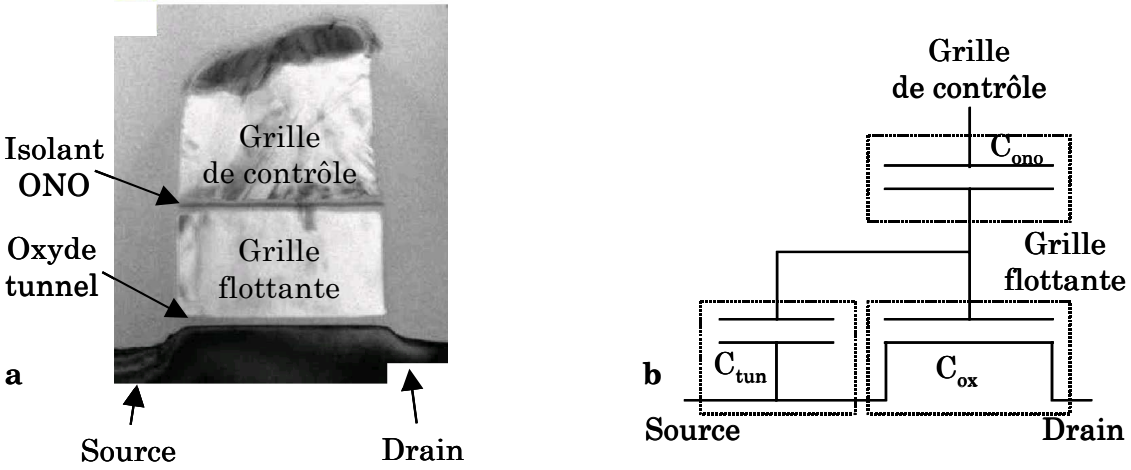

Figure I.12. *Coupe SEM (Scanning Electron Microscopy) d'une mémoire de type Flash (***a***) et schéma électrique équivalent faisant apparaître les différentes capacités (***b***), [Laffont'03b].*

Les figures (I.12.a) et (I.12.b) montrent une coupe SEM (Scanning Electron Microscopy) et le schéma électrique équivalent d'une mémoire Flash. On peut identifier les trois composants principaux que sont : le transistor MOS (avec sa capacité C_{ox}), la capacité inter-poly (C_{ono}) et la capacité de recouvrement de la source (C_{tun}).

I.4.2.2. Architecture des mémoires Flash

Les mémoires Flash peuvent être regroupées en une architecture de type NAND ou de type NOR [Cappelletti'99]. Quelle que soit l'architecture le plan mémoire est constitué d'une matrice de lignes (Word line) et de colonnes (bit line) dont l'intersection correspond à un point mémoire. La figure (I.13.a) présente l'architecture de type NOR. Durant les opérations de lecture, la cellule lue est adressée en polarisant sa word line positivement alors que les autres word lines sont connectées à la masse. Afin d'éviter toute perturbation de la bit line par des courants de fuites, les cellules non sélectionnées doivent avoir une tension de seuil équivalente positive. Le principal inconvénient de ces architectures NOR est leur faible densité d'intégration puisque tous les points mémoires ont leur drain connecté à la bit line et leur source à la ligne commune.

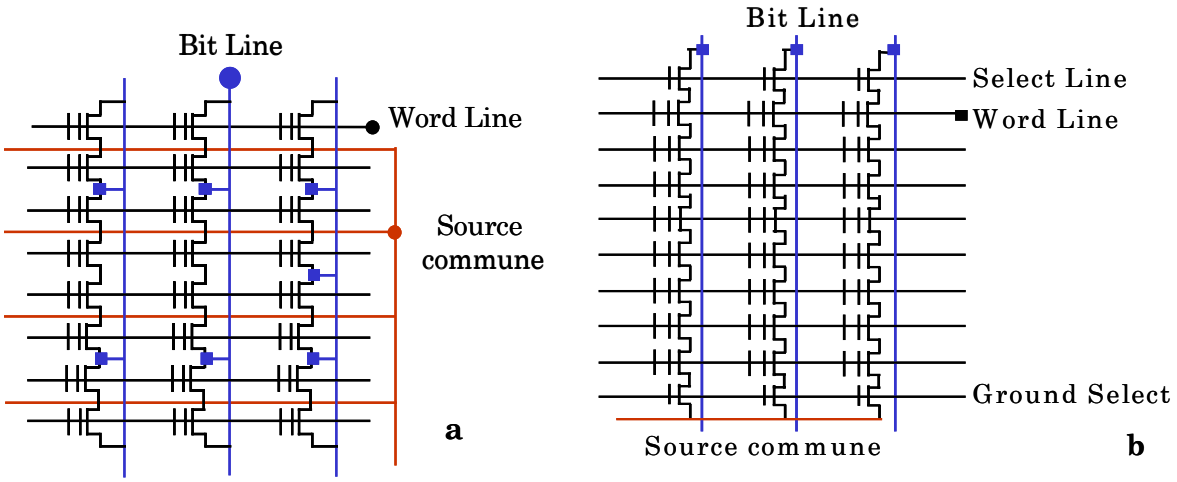

Figure I.13. *Architecture de type NOR* (**a**) *et de type NAND* (**b**).

Une meilleure densité est obtenue grâce à l'architecture NAND présentée dans la figure (I.13.b). Dans ce cas, les bit lines sont des lignes composées de points mémoires connectés en série. Deux transistors de sélection sont situés sur chaque ligne: le premier, qui sert à sélectionner la bit line, est commandé par le signal SL (« Select Line ») et le second, qui sert à relier les bit lines à la masse, est commandé par le signal GS (« Ground Select »). Afin de choisir une cellule de la ligne, sa word line doit être activée, ainsi que toutes les word lines commandant les autres cellules de la ligne. Pour lire l'état du point mémoire sélectionné, une tension de lecture assez faible est appliquée sur sa grille, alors

qu'une tension supérieure à la tension de seuil équivalente maximale est appliquée aux autres points mémoire. Ainsi la cellule sélectionnée impose le courant de la bit line à lire.

I.4.2.3. Principe de fonctionnement des mémoires Flash

Dans ce paragraphe, nous décrirons uniquement le fonctionnement d'un seul point mémoire. Comme pour tous les dispositifs MOS à grille flottante, le MOSFET fonctionne comme un interrupteur avec une modulation des électrons du canal par la grille de contrôle (GC pour Gate Control). La grille flottante, déconnectée des électrodes où sont appliquées les tensions, joue le rôle d'élément mémoire. Ainsi, la caractéristique I_{DS} (V_{GC}) d'une structure à grille flottante dépend de la charge stockée dans celle-ci, Q_{FG}, qui induit une variation de la tension de seuil entre deux valeurs distinctes (cf. Fig. (I.14)). Soit V_{T1}, la tension de seuil initiale du dispositif. L'état écrit de la mémoire résulte du stockage d'électrons dans la grille flottante. La tension de seuil du MOSFET augmente et atteint une valeur V_{T2}, le transistor est alors bloqué. Pour effacer la mémoire, les électrons sont évacués de la grille flottante et la tension de seuil retrouve sa valeur initiale, V_{T1}. Le transistor est alors passant. Notons que la différence, ΔV_T, entre les tensions de seuil de l'état écrit et de l'état effacé correspond à la fenêtre de programmation de la mémoire. L'état de la mémoire est déterminé par une mesure en courant du transistor en polarisant la grille de contrôle par une tension appartenant à la fenêtre de programmation de la mémoire.

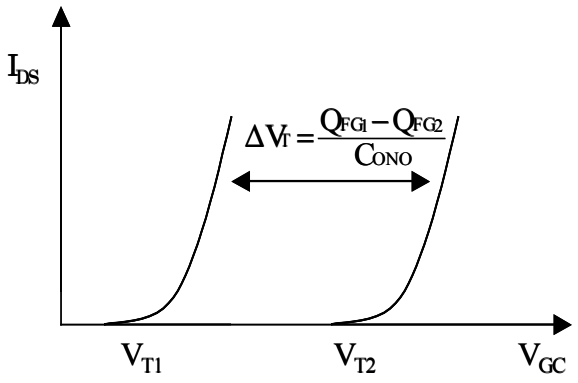

Figure I.14. *Caractéristiques I_{DS} (V_{GC}) d'une structure à grille flottante pour deux charges différentes sur la grille flottante.*

Pour une cellule flash, l'effacement est obtenu par injection d'électrons de la grille flottante vers la source (et le substrat) par le biais d'un courant Fowler-Nordheim (I_{FN}). L'écriture peut être obtenue par injection FN (cas des architectures NAND) ou par injection par porteurs chauds (cas des architectures NOR). Dans ce dernier cas, le transistor est polarisé en régime de saturation. Il existe alors à la jonction canal/drain polarisée en inverse, un champ électrique d'autant plus important que la longueur du canal diminue ($\xi = V / L$). Par conséquent, les électrons qui pénètrent dans la zone de

désertion sont accélérés par ce champ électrique très intense (forte courbure des bandes d'énergie). Ce phénomène est illustré sur la figure (I.15) par le repère (1). Certains électrons acquièrent alors suffisamment d'énergie cinétique (porteurs chauds) pour se comporter comme des particules ionisantes (collision avec le réseau cristallin) et générer des paires électrons-trous : c'est le phénomène d'ionisation par impact. On obtient donc deux électrons dans la bande de conduction et un trou dans la bande de valence. Les paires électrons-trous ainsi créées sont dissociées sous l'effet du champ électrique. Les trous peuvent être attirés par l'électrode de substrat et donner naissance à un important courant de substrat I_{SUB} (repère (2)). Ils peuvent également migrer vers la source et créer un abaissement de la barrière à la jonction source/canal. Il se produit alors une injection d'électrons supplémentaires de la source vers le canal. En toute rigueur, le courant de substrat (I_B) est la somme de I_{SUB} et du courant des jonctions source/substrat et drain/substrat ($I_B = I_{SUB} + I_{diode}$). Pour nos travaux, le courant inverse des diodes sera toujours négligé ($I_B \approx I_{SUB}$). La majeure partie des électrons générés, par ionisation par impact, s'additionne au courant du transistor I_{DS} pour donner le courant noté I_D (repère (3)). Nous en déduisons qu'en présence de porteurs chauds, le courant de drain n'est plus égal au courant de source qui lui, est toujours égal à I_{DS} ($I_S = I_{DS}$). Enfin, lorsque les électrons sont accélérés au niveau de la jonction canal/drain, une polarisation positive de la grille donne la possibilité à ceux qui ont acquis une énergie potentielle suffisante de franchir la barrière de potentiel de l'interface silicium/oxyde, ce qui correspond au courant de grille, I_G (repère (4)).

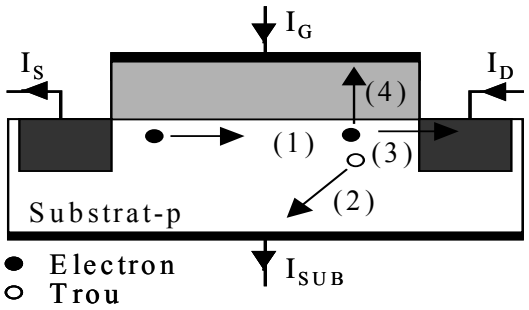

Figure I.15. *Localisation de la génération des paires électron-trou due à la présence d'électrons chauds près du drain. Les trous sont collectés par la prise substrat (I_{SUB}). Une partie des électrons générés s'additionne au courant de drain alors que le reste traverse l'isolant de grille (I_G).*

Afin de simplifier l'étude, on considèrera seulement trois conditions élémentaires pour l'injection de porteurs chauds vers la grille :

- Les porteurs doivent posséder une énergie suffisante afin de franchir la barrière de potentiel oxyde / semiconducteur.
- Les porteurs doivent avoir une direction perpendiculaire à l'interface Si/SiO_2.

- Lors de leur parcours dans le réseau, après avoir acquis une énergie suffisante, les porteurs ne doivent pas avoir d'interaction avec le réseau et ainsi conserver leur énergie, il en va de même dans l'oxyde.

Pour nos travaux de thèse, nous nous somme basés sur l'équation du courant d'injection de porteurs chauds (CHEI pour Channel Hot Electron Injection) donnée par Tam *et al.* [Tam'84] :

$$I_G = I_{sub}\,\alpha_{ox}\,\exp\left(-\frac{b_{ox}}{\xi_{ox}}\right) \qquad \text{(I.47)}$$

où b_{ox} et α_{ox} sont les deux paramètres d'injection.

Dans cette équation b_{ox} est le facteur de dépendance du courant de grille avec le champ électrique dans l'oxyde tunnel (paramètre représentant la probabilité de passage d'un électron à travers l'interface Si/SiO_2 des modèles classiques et de l'électron chanceux) et α_{ox} celui du courant grille par rapport au courant substrat. Le courant substrat a pour expression d'après le modèle de Schokley Read Hall :

$$I_{sub} = I_{DS}\,\frac{ai}{bi}\left(V_D - V_{D\,sat}\right)\exp\left(\frac{-bi}{V_D - V_{D\,sat}}\right) \qquad \text{(I.48)}$$

où ai, bi sont les coefficients d'ionisation par impact et V_{sat} le potentiel appliqué aux bornes de la zone à saturation.

Les coefficients ai, bi, α_{ox} et b_{ox} sont obtenus à partir des caractéristiques statiques $I_{SUB}(V_{GS},V_{DS})$ et $I_G(V_{GS},V_{DS})$ mesurées sur des mémoires dont la grille flottante est reliée à la grille de contrôle (aussi appelées dummy cell).

En raison de la grande intensité de ce dernier type d'injection, l'opération d'écriture d'une mémoire Flash est extrêmement rapide comparée à l'opération d'effacement (courant FN). Cette particularité rend la mémoire Flash très attractive par rapport à la mémoire EEPROM. Cependant, la limitation de la surface consacrée à la mémoire et le volume croissant du stockage souhaité nécessite la réduction de la taille des composants. Suivant la Roadmap International Technologie Roadmap for Semiconductor (ITRS) 2003, la taille limite des mémoires flash serait de 65 nm avec une épaisseur d'oxyde tunnel de l'ordre de 8-9 nm. Cette taille critique est due à l'incompatibilité entre la réduction de l'épaisseur de l'oxyde de grille pour contrôler les phénomènes de canaux courts et la préservation d'une épaisseur de diélectrique minimum pour maintenir sa fiabilité et la rétention de la charge après plusieurs cycles d'écriture et d'effacement. De plus, dès 1990, Bez *et al.* [Bez'90] ont mis en évidence la limitation de la réduction de la longueur

des mémoires flash placées dans une architecture NOR à cause du phénomène appelé « Drain turn on », engendré par le fort couplage entre le drain et la grille flottante. Ce phénomène se traduit par le contrôle du canal par la polarisation de drain lorsque la grille n'est pas (ou peu) polarisée. La discrétisation de la grille supprime le couplage entre le drain et la grille flottante [Lombardo'04] ce qui induit la réduction de l'influence des phénomènes de canaux courts, et permet l'utilisation de tensions de drain plus élevées pour l'opération de lecture.

Ainsi, l'utilisation de stockages discrets en remplacement des traditionnels stockages continus dans la grille flottante est une des solutions envisagées pour surmonter la limitation de dimensions. Il existe plusieurs types de mémoires à piégeages discrets décrits dans la littérature car le matériau utilisé pour le stockage peut être du nitrure présentant de nombreux défauts naturels, ou des matériaux High K (Al_2O_3 ou HfO_2) ou des nano-cristaux de semiconducteur [Tiwari'95] et [Shi'98].

I.4.3. Les mémoires à nano-cristaux

Depuis les années 1990, les nano-cristaux (ou nodules, ou encore dots) de silicium sont très étudiés pour leurs propriétés physiques mais également pour la fabrication de nouveaux dispositifs pour la microélectronique et la photonique. En 1995, Tiwari utilise des nano-cristaux de silicium à la place des traditionnelles grilles flottantes [Tiwari'95]. Ces dispositifs sont très prometteurs en terme de réduction de dimensions car ils présentent l'avantage d'une haute densité d'intégration, d'une basse consommation en puissance (environ 12V contre 18-20V pour les mémoires Flash traditionnelles) et d'un bas coût de fabrication puisque les nodules ne nécessitent pas d'étape de masquage pour les isoler électriquement. Ainsi, le nombre de masques nécessaire à la fabrication des dispositifs décroît de 11 masques pour les Flash à grille flottante traditionnelle à 4 masques pour les mémoires à nano-cristaux [Chang'03].

I.4.3.1. La structure des mémoires à nano-cristaux de silicium

La figure (I.16) présente une coupe TEM (Transmission Electron Microscopy) d'une mémoire à nodules de diamètre 5 nm et de longueur, L = 0.2 à 0.3 µm. Cette mémoire est un transistor d'apparence classique si ce n'est la présence de "boules" ou de "demi-boules" de silicium, de tailles nanométriques, réparties dans l'oxyde de grille à une certaine distance de l'interface, recouvrant entièrement la surface du canal. La densité de nodules est de l'ordre de 10^{12} dots.cm^{-2} et l'épaisseur de la couche de diélectrique

séparant les nano-cristaux du substrat est contrôlée afin de diminuer la dispersion de la distribution des tensions de seuils.

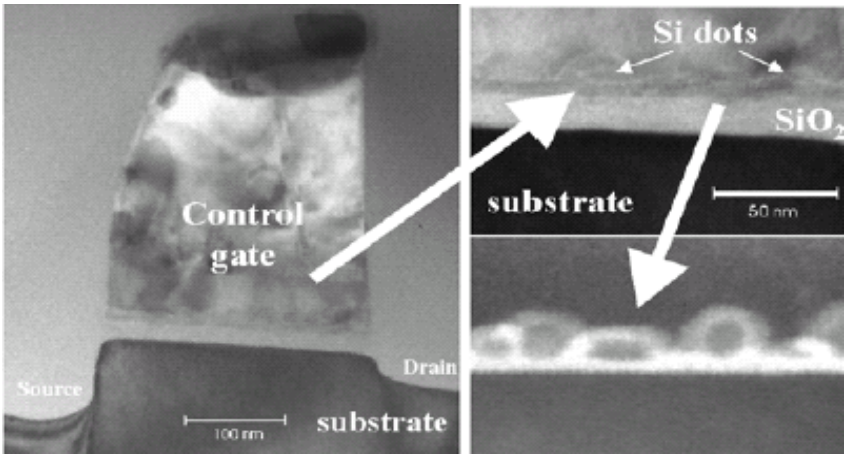

Figure I.16. *Coupe TEM (Transmission Electron Microscopy) d'une mémoire à nodules de diamètre 5nm et de longueur L = 0.2 à 0.3 µm [Corso'03].*

D'autres paramètres comme la taille des nano-cristaux, leur forme, l'isolation latérale entre les nano-cristaux, et l'uniformité de la densité surfacique des nano-cristaux doivent également être contrôlés lors du procédé de fabrication pour améliorer les performances de ces mémoires. Ainsi plusieurs techniques de fabrication de nano-cristaux ont successivement été envisagées : la croissance des nano-cristaux par dépôt LPCVD (Low pressure chemical vapor deposition) [Tiwary'95], la précipitation du silicium avec implantation ionique [Hanafi'96], le dépôt par aérosol [Debauwe'00]. Très récemment, une nouvelle technique de dépôt en deux temps, à partir de deux gaz différents (SiH_4 et SiH_2Cl_2), a mené à la séparation des phases de nucléation et de croissance des nano-cristaux, permettant une meilleure maîtrise de leur densité, de leur taille et par conséquent de leur isolement [De Salvo'03].

Hormis les mémoires à nano-cristaux présentées par Tiwary, il existe d'autres types de mémoires à nano-cristaux selon les matériaux utilisés (pour les nodules ou pour le diélectrique de grille) et la disposition des nano-cristaux. En 1998, une alternative aux nano-cristaux de silicium est proposée par l'Université de Berkeley avec les nano-cristaux de germanium. Ces mémoires présentaient de meilleures caractéristiques d'écriture et d'effacement et de meilleurs temps de rétentions que celles des dispositifs à nano-cristaux de silicium [King'98]. En 1998, Kim *et al.* [Kim'98] ont présenté également des mémoires à nano-cristaux de silicium utilisant un diélectrique formé d'oxyde nitrure permettant une meilleure uniformité dans la répartition des dots (à cause de la rugosité de la surface du nitrure). Puis, en 2002, les mémoires à nodules de métal ont été

également proposées [Liu'02]. Ces dispositifs présentent une plus forte densité d'états autour du niveau de Fermi (c.a.d. une plus grande protection contre la fluctuation des niveaux de Fermi causée par des contaminations), une plus grande gamme de valeur de travail de sortie et de plus petites perturbations d'énergie dues au confinement des porteurs.

Dans la suite de ce manuscrit, nous nous limiterons aux dispositifs à nodules de silicium.

I.4.3.2. Fonctionnement des mémoires à nano-cristaux de silicium

La fonction mémoire de ces dispositifs est attribuée à l'échange de charges entre les nano-cristaux de silicium et la couche d'inversion à travers un diélectrique tunnel fin, t_{ox1}. L'isolement des nodules les uns des autres [Chae'99], empêche le mouvement latéral des charges et préserve la mémoire d'une perte totale de l'information lors d'une détérioration locale de l'oxyde. En effet, les mémoires à nano-cristaux de silicium sont des mémoires à stockage discret pour lesquelles quelques électrons sont stockés dans chaque nodule (selon la taille de ces derniers). La charge emmagasinée dans l'ensemble des nodules contrôle la conductivité du canal du transistor mémoire.

L'injection d'un électron à partir de la couche d'inversion s'effectue par effet tunnel lorsque la grille est en polarisation directe par rapport à la source et au drain. La charge stockée écrante la charge de la grille et réduit la conduction dans la couche d'inversion, et par suite entraîne une augmentation de la tension de seuil. Le chargement de ces nodules, avec des électrons, peut se faire par injection Fowler-Nordheim en appliquant une tension de grille positive ou par porteurs chauds en appliquant une tension positive sur la grille et sur le drain et/ou sur la source.

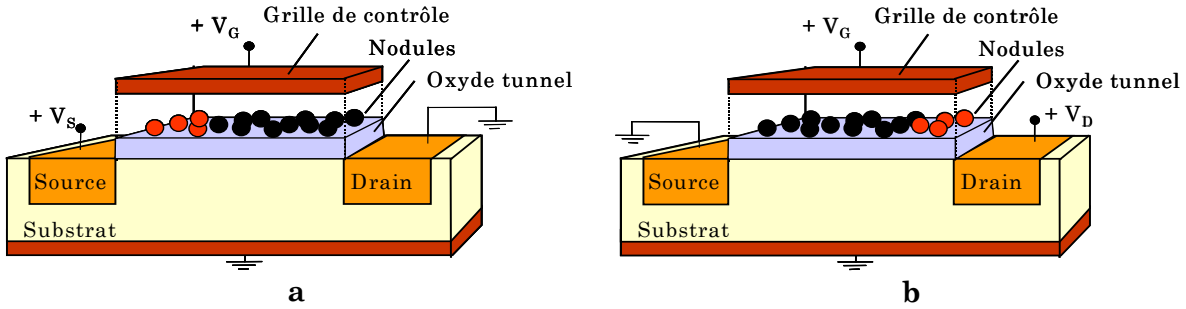

Figure I.17. *Schématisation du fonctionnement 2 bits d'une mémoire à nodules avec écriture côté source (***a***) et côté drain (***b***).*

Par conséquent, l'écriture par porteurs chauds localise l'injection des porteurs dans la région du canal proche du drain et/ou de la source suivant la polarisation choisie (tension

positive sur le drain et/ou la source (cf. Fig. (I.17)). Ainsi, les mémoires à nodules présentent la possibilité d'être utilisées comme des mémoires 2 bits [Eitan'99], [Eitan'00] et [Hradsky'03]. Ce concept de « dual bit » a récemment été amélioré avec des dispositifs à nano-cristaux de métal comportant deux sources et deux drains ce qui permet l'obtention de 4 bits par cellule mémoire [Liu'03].

Le déchargement des nodules (effacement de la mémoire) s'effectue par courant Fowler-Nordheim en appliquant un potentiel de grille négatif. Comme pour la mémoire Flash l'état écrit ou effacé de la mémoire se fera par la détermination de la tension de seuil au cours d'une lecture $I_{DS}(V_{GS})$.

De part leurs petites dimensions, leur fonctionnement à basses tensions, les temps d'écriture et d'effacement (respectivement de l'ordre de la micro-seconde et de la mili-seconde) et leur endurance (10^5 cycles écriture/effacement), les mémoires à nodules sont de bons candidats pour les applications spatiales [Bell'01] et les applications commerciales de type téléphones portables ou ordinateurs portables. Par exemple, grâce à leur fort potentiel de miniaturisation, ces mémoires peuvent être utilisées pour des stockages de photos de caméras digitales. Enfin, notons que les dispositifs à nano-cristaux de silicium ont également des propriétés photoniques. En effet, en 2000, Patch [Patch'00], a montré que des nano-cristaux de silicium emprisonnés dans une couche de SiO_2 pouvaient émettre de la lumière lors de stimulations électriques.

I.5. Conclusion

L'objectif de ce premier chapitre était d'introduire les différentes notations utilisées dans la suite du manuscrit. Nous avons ainsi pu rappeler les principales caractéristiques et le mode de fonctionnement des différentes structures que nous avons étudiées durant nos travaux de thèse, à savoir les capacités MIS, les transistors MOS et les mémoires Flash à nodules de silicium. Les relations de base sur lesquelles reposent nos modèles ont été présentées. En ce qui concerne la modélisation du courant de drain du transistor MOS en inversion faible, et en inversion forte avant saturation, nous utiliserons le modèle de Pao et Sah [Pao'66] ou le modèle en feuillet [Brews'78] basés sur le calcul du potentiel de surface. De plus, les structures étudiées ayant des isolants de grille minces, nous avons également décrit les modes d'injections tunnel (Tunnel Direct et Fowler-Nordheim), ainsi que l'injection par porteurs chauds utilisée pour l'écriture des mémoires Flash à nodules.