Etude des structures à nano-cristaux de silicium

IV.1. Introduction

Au cours de ces dix dernières années, les nano-cristaux de silicium ont été introduits dans les mémoires Flash afin d'améliorer leurs performances : réduction de dimensions, robustesse par rapport aux défauts de l'oxyde, basse consommation en puissance, bas coût de fabrication, etc… L'objectif de ce chapitre est de présenter une modélisation du comportement électrique des mémoires à nodules de silicium lors de leur écriture en fonction des paramètres des dots. Dans un premier temps, nous avons étudié l'effet de la répartition non uniforme de la charge stockée sur les caractéristiques électriques de ces dispositifs, puis nous avons développé un modèle pseudo 2D de l'opération d'écriture de la mémoire par porteurs chauds. La dernière partie de ce chapitre est consacrée à la caractérisation électrique de structures à nodules fabriquées par la société STMicroelectronics Catagne dans le cadre du projet européen ADAMANT (Advanced Memories based on Discrete Traps).

IV.2. Structures d'étude

Les nodules considérés dans notre étude sont des nano-cristaux de silicium séparés les uns des autres par du SiO_2. Ils remplacent la grille flottante conventionnelle des mémoires Flash et sont constitués de demi-sphères en silicium offrant leur section la plus large à l'injection tunnel (cf. Fig. (IV.1)).

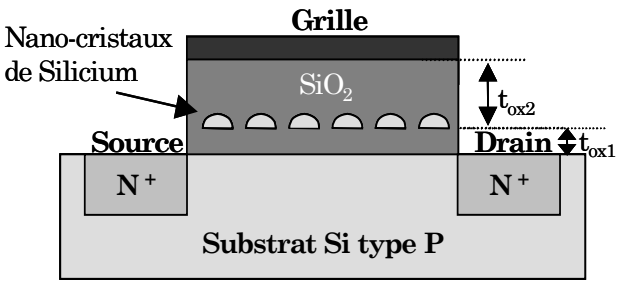

Figure IV.1. *Coupe schématique du dispositif mémoire Flash comportant des nodules de silicium.*

Ces îlots sont supposés suffisamment espacés les uns des autres, pour empêcher le mouvement latéral des charges et préserver la mémoire d'une perte totale de

l'information lors d'une détérioration locale de l'oxyde. De surcroît, ces nodules sont supposés uniformément répartis dans l'oxyde du transistor entre l'oxyde tunnel d'épaisseur (t_{ox1}) et l'oxyde de contrôle d'épaisseur (t_{ox2}). Comme le montre la figure (IV.1), seule une faible portion R_{eff} de la surface totale de l'oxyde de grille est occupée par les nano-cristaux de silicium. Ce coefficient R_{eff} est proportionnel à la densité (N_{dot}) et au diamètre (D_{dot}) des nodules :

$$R_{eff} = N_{dot}\,\pi\left(\frac{D_{dot}}{2}\right)^2 \qquad \text{(IV.1)}$$

Comme nous l'avons rappelé dans le chapitre I, l'opération d'écriture des mémoires Flash à nodules consiste à stocker des électrons dans les nano-cristaux, soit par le biais du mécanisme Fowler–Nordheim (FN), I_{FN}, qui est assez lent (de l'ordre de la micro-seconde) et uniforme (cf. Fig. (IV.2.a)), soit par le mécanisme de porteurs chauds (CHE) [Tam'84], qui est plus rapide (facteur 10) mais plus coûteux en terme d'énergie car le transistor doit être polarisé en régime de saturation (V_{DS} fort) pour créer des paires électrons-trous (par ionisation par impact). Toutefois, l'écriture par porteurs chauds permet un accès sélectif à un seul point mémoire (cf. Fig. (IV.2.b)) puisque la création de paires électrons-trous est localisée dans une zone proche du drain ou de la source suivant la polarisation (V_D-V_S) appliquée [Eitan'00]-[Lusky'01]. Les trous ainsi générés se dirigent vers le substrat, tandis que les électrons suivent différents chemins : sous l'influence du champ électrique entre l'interface et la grille, une partie des électrons traverse l'isolant de grille du transistor pour atteindre soit la grille, soit les nodules (qu'ils chargent) ; la majeure partie des électrons générés par ionisation rejoignent le drain (ou la source suivant leur point de départ), sous l'influence du champ électrique entre les extrémités du canal.

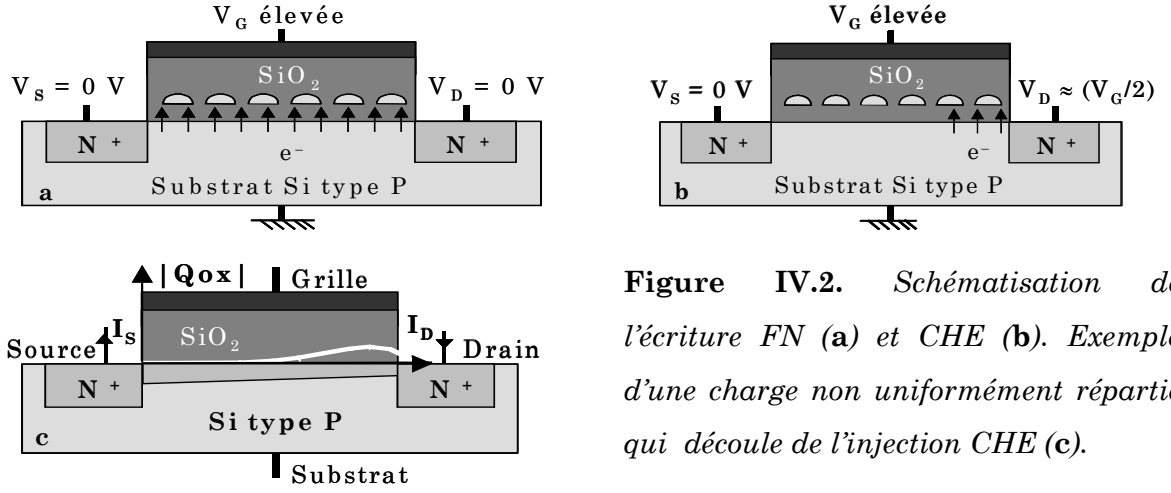

Figure IV.2. *Schématisation de l'écriture FN (***a***) et CHE (***b***). Exemple d'une charge non uniformément répartie qui découle de l'injection CHE (***c***).*

Par exemple, la figure (IV.2.c) présente une schématisation de la non uniformité de charges qui découlerait de charges stockées uniquement à l'intérieur de quelques nodules proches du drain. Cette possibilité de chargement discret, localisé près d'une jonction (drain et/ou source) offre la possibilité d'une logique 2 bits pour les mémoires à nodules [Bloom'01], c'est à dire quatre états possibles :

- aucun nodule chargé,
- tous les nodules chargés par injection Fowler-Nordheim (charge uniformément répartie dans les nodules),
- des charges présentes dans quelques nodules proches de la source (injection localisée par porteurs chauds).
- des charges présentes dans quelques nodules proches du drain (injection localisée par porteurs chauds).

IV.3. Modélisation d'une charge non uniformément répartie

IV.3.1. Modélisation pour un MOSFET

Considérons les profils de charges en exponentielle décroissante tracés sur la figure (IV.3.a). La charge stockée au dessus du canal modifie la conductivité du canal du transistor par l'intermédiaire de la variation de la tension de bandes plates, V_{FB} (cf. Fig. (IV.3.b)).

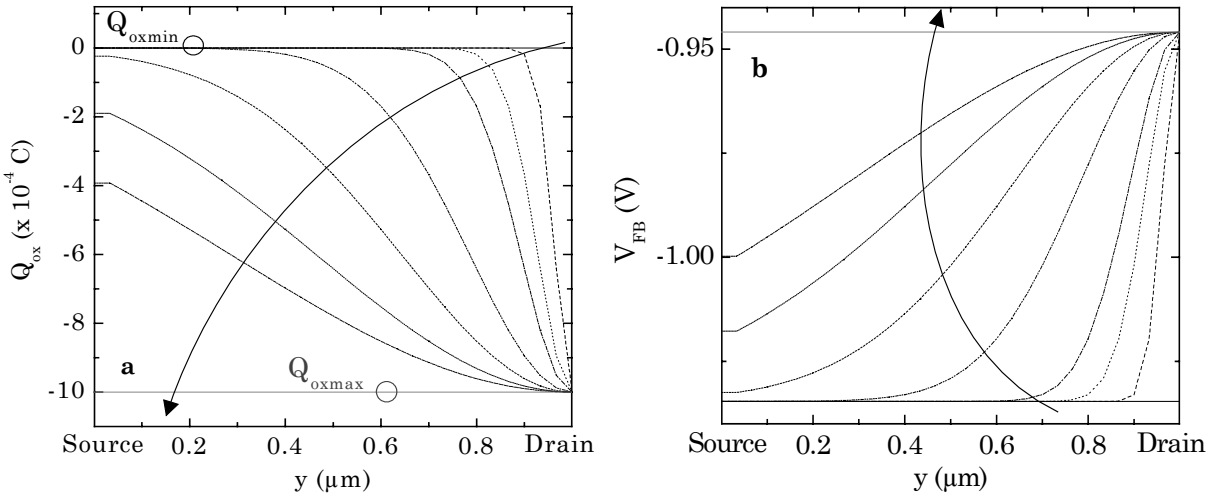

Figure IV.3. *Profils de charges non uniformément réparties dans l'oxyde (***a***) et variations des tensions de bandes plates correspondantes (***b***).*

Comme nous l'avons rappelé au chapitre III, les modélisations de type Pao et Sah et en feuillet ne prennent pas en compte ce type de non uniformité dans le développement de l'expression du courant de drain. Le modèle segmenté, développé lors de nos travaux sur les non uniformités des transistors (voir chapitre III), permet de surmonter cette difficulté. Par conséquent, pour simuler le courant de drain des transistors ayant les profils de charges reportés sur la figure (IV.3.a), nous avons adapté le modèle segmenté en considérant la structure comme équivalente à la juxtaposition de N transistors (de longueur L/N) ayant chacun une charge fixe constante dans l'oxyde mais pouvant être différente d'un segment à l'autre. Le courant de drain des N transistors élémentaires est évalué à l'aide du modèle en feuillet [Brews'78]. A des tensions de grille (V_{GS}) et de drain (V_{DS}) fixées, le potentiel de surface (et par conséquent l'écart entre les quasi-niveaux de Fermi, Φ_C), est calculé pour chaque transistor en supposant un flux de courant conservatif le long du canal. Le système de N équations à N–1 inconnues peut alors être résolu par la méthode du pont diviseur. Cependant, comme nous l'avons déjà signalé, l'utilisation du pont diviseur n'est valide que si la somme de tous les courants tunnels est négligeable par rapport au courant de drain I_{DS}. Pour les structures considérées, ces conditions sont respectées puisque le courant d'injection est très petit devant I_{DS} et que les fuites de grille sont considérées comme négligeables.

Les figures (IV.4) montrent le décalage de la courbe $I_{DS}(V_{GS})$ d'un transistor NMOS obtenu pour une tension de drain V_D = 50 mV (avec V_S = 0 V). Ce décalage résulte de la présence de charges, Q_{ox}, qui correspondent aux profils présentés à la figure (IV.3.a).

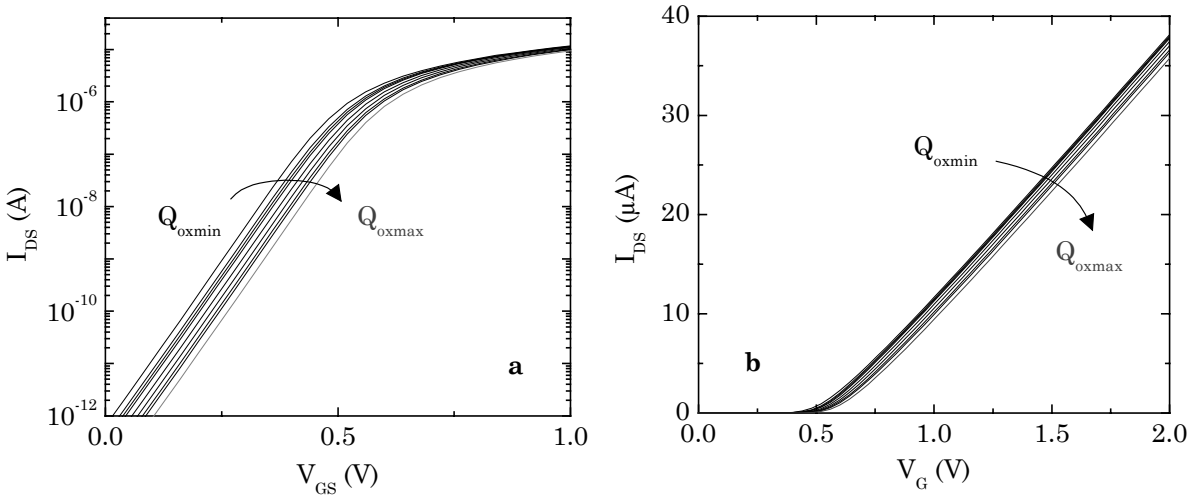

Figure IV.4. *Variations du courant de drain en présence de charges non uniformément réparties (cf. Fig. (IV.3.a)) tracées en échelle semi-logarithmique (***a***) et en échelle linéaire (***b***). Les paramètres du MOSFET sont : W = 1 µm, L = 1 µm, t_{ox1} = 3 nm et V_{DS} = 50 mV.*

Comme pour une variation du dopage de substrat entre le drain et la source, le décalage des caractéristiques $I_{DS}(V_{GS})$ est continu entre la courbe correspondant à une charge nulle et celle correspondant à une charge maximale.

IV.3.2. Modélisation d'une mémoire à nodules

Dans ce paragraphe, nous ne considèrerons que l'écriture, par porteurs chauds, des mémoires Flash à nodules, puisque ce mécanisme permet une injection localisée des électrons à partir du canal.

Pour déterminer la charge stockée dans les nodules, la valeur de Φ_C doit être calculée le long du canal. Comme pour un transistor, la mémoire à nodules peut-être supposée équivalente à N transistors juxtaposés de longueur L/N (cf. Fig. (IV.5)).

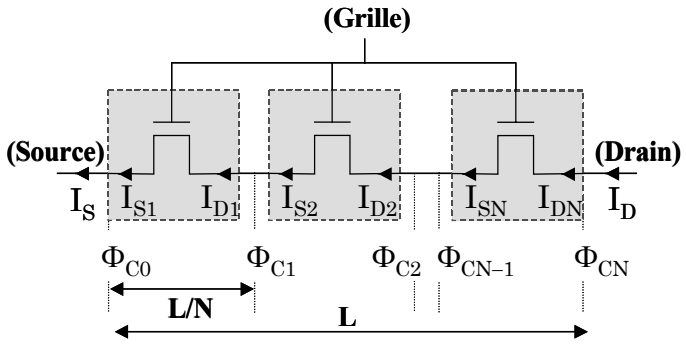

Figure IV.5. *Vue schématique de la mémoire à nodules découpée en N transistors.*

Ainsi, ce modèle permet la modulation de la longueur de chaque transistor et plus particulièrement sa réduction dans la zone d'injection (près du drain de la mémoire). Notons que cette approche offre d'autres possibilités comme la simulation du courant de drain pour des mémoires ayant des charges stockées à la fois près du drain et de la source (mémoires 2 bits [Eitan'99],[Eitan'00]).

La charge totale Q_{dot} (exprimée en Coulombs) étant stockée dans les nodules enfermés dans la couche d'oxyde d'épaisseur t_{ox2}, le potentiel de surface aux frontières de chaque transistor est alors évaluable à partir de l'équation suivante :

$$V_{GB} = \left(V_{FB} + \Psi_S - \frac{Q_{SC}}{C_{ox}} \right) - \frac{1}{C_{ox}} \frac{Q_{dot}}{W(L/N)} \frac{t_{ox2}}{t_{ox1} + t_{ox2}} \qquad \text{(IV.2)}$$

où W est la largeur de la mémoire.

Cette équation (IV.2) suggère que l'influence électrostatique des charges stockées dans les nodules est étalée sur la surface totale de chaque transistor élémentaire (comme pour la modélisation des pièges dans un diélectrique).

Le chargement des îlots de silicium s'effectuant par le biais de porteurs chauds, qui apparaissent près du drain lorsque le transistor MOS est polarisé en régime saturé (forte tension V_{DS}), il est nécessaire d'évaluer le champ électrique dans la zone saturée, le courant d'ionisation par impact et par suite, le courant qui traverse l'isolant tunnel.

Usuellement, le canal est dit pincé lorsque la charge d'inversion devient pratiquement nulle au niveau du drain. Cependant, pour une tension de grille donnée, la localisation du point de pincement dépend de la tension V_{DS}. En effet, une augmentation de la polarisation appliquée sur le drain entraîne une augmentation de la largeur de la zone de charge d'espace (ZCE) de la jonction drain/substrat vers le canal (qui est moins dopé que le drain). Par suite, lorsque le potentiel de saturation V_{DSsat} est dépassé l'accroissement de la ZCE conduit à un déplacement du point de pincement vers la source. Dans notre modèle, nous supposons que le point de pincement est atteint lorsque la charge d'inversion dans le canal devient négligeable devant la charge d'inversion de la source [Laffont'03a]. Ainsi, pour des tensions de drain et de grille données, la zone de saturation est localisable le long du canal en recherchant le quasi-niveau de Fermi Φ_{Csat} tel que :

$$Q_n\left(\Phi_{Csat}, V_{GB}\right) = \frac{Q_n\left(V_{SB}, V_{GB}\right)}{FAC} \qquad \text{(IV.3)}$$

où FAC est un paramètre d'ajustage (typiquement égal à 10).

Par conséquent, le potentiel de saturation V_{DSat} est obtenu en fonction de la résolution spatiale de Φ_C et permet de connaître à la fois la localisation (dans la zone de pincement) et le nombre (N_{pinch}) de transistors élémentaires dont les dots se remplissent par porteurs chauds.

Comme nous l'avons rappelé dans le chapitre I, le courant injecté, I_{CHE}, est fonction du courant de substrat I_{sub}, provenant de l'ionisation par impacts exprimée par l'expression (I.49) dans le modèle de Tam [Tam'84]. Toutefois, contrairement aux mémoires à grille flottante conventionnelle, seule une partie de la charge injectée peut être stockée puisque les nodules ne recouvrent qu'une portion de la surface de l'interface entre les couches d'oxyde t_{ox1} et t_{ox2}. Par suite, le coefficient (R_{eff}) représentant la surface occupée par les nodules a été introduit dans l'expression du courant injecté. En conséquence, en supposant que l'injection soit uniforme dans la zone à saturation pour chaque transistor localisé dans la région de pincement, le courant réellement injecté est donné par la relation suivante :

$$I_W = I_{sub}\alpha_{ox}\frac{R_{eff}}{N_{pinch}}\exp\left(\frac{b_{ox}}{\xi_{oxt1}}\right) \quad \text{(IV.4)}$$

où b_{ox} et α_{ox} sont les deux paramètres d'injection et I_{sub} représente le courant substrat décrit par le modèle de Schokley Read Hall :

$$I_{sub} = I_{DS}\frac{ai}{bi}V_{sat}\exp\left(-\frac{bi}{V_{sat}}\right) \quad \text{(IV.5)}$$

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

Notons que pour des mémoires Flash conventionnelle, les coefficients ai, bi, α_{ox} et b_{ox} sont déterminés à partir des caractéristiques statiques $I_{SUB}(V_{GS},V_{DS})$ et $I_G(V_{GS},V_{DS})$ mesurées sur une cellule dont la grille de contrôle et la grille flottante sont reliées entre elles.

Le paramètre ξ_{oxt1} correspondant au champ électrique dans le diélectrique tunnel, entre l'interface et les dots, est donné par l'expression suivante :

$$\xi_{oxt1} = \frac{-V_{ox}}{t_{ox1} + t_{ox2}} - \frac{t_{ox2}}{t_{ox1} + t_{ox2}}\frac{Q_{dot}}{\varepsilon_{ox}W(L/N)R_{eff}} \quad \text{(IV.6)}$$

où V_{ox} est le potentiel diélectrique total et Q_{dot} résulte de l'intégration du courant d'écriture :

$$\Delta Q_{dot} = I_W \times \Delta t \quad \text{(IV.7)}$$

où Δt, correspond au pas du temps d'écriture.

Remarquons que, pour les mémoires Flash à nodules, à une tension de grille donnée, l'injection dans les nodules est stoppée lorsque le champ électrique, ξ_{oxt1}, tend vers zéro, même si le courant par porteurs chauds I_{CHE} existe toujours. Dans ce cas, la barrière de potentiel n'est plus assez déformée pour laisser passer les électrons par effet tunnel FN. Cette notion est très différente de celle des mémoires Flash traditionnelles pour lesquelles l'opération d'écriture s'arrête lorsque le courant I_{CHE} devient négligeable (le nombre d'électrons injectés entraîne la « dé-saturation » du transistor). De même, lors de l'effacement de la mémoire à nodules, le champ à considérer dans l'expression du courant correspond à ξ_{oxt1}.

IV.4. Simulations des structures 1bit

Afin de simplifier notre modélisation pour une étude 1 Bit, nous avons supposé que la mémoire à nodules était équivalente à un transistor coupé en deux parties : l'une proche

du drain, de longueur X_D, correspondant à la région de chargement et l'autre près de la source de longueur L – X_D, correspondant à la partie non chargée du transistor (cf. Fig. (IV.6)).

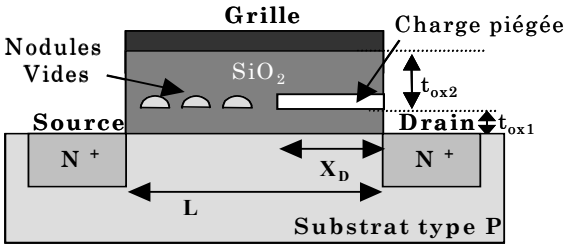

Figure IV.6. *Coupe schématique du dispositif mémoire Flash comportant deux parties : une chargée et une non chargée.*

Notons que lors de cette étude du chargement des nodules, les effets de canaux courts ne sont pas pris en compte.

IV.4.1. Simulations statiques des mémoires Flash à nodules

La figure (IV.7) montre les courbes $I_{DS}(V_{GS})$ simulées pour un transistor NMOS. Le décalage observé est dû à la présence d'une forte densité de charges, N_Q, localisée dans l'isolant (à 5 nm de l'interface) sur une longueur X_D proche du drain. Pour V_{GS} donnée, cette forte densité de charges induit un fort décalage de la pente sous le seuil pour les transistors chargés sur une petite zone ($X_D < 10^{-2}$ L) par rapport au transistor sans charge (c.a.d. $X_D = 0$) (cf. Fig. (IV.7.b)). Pour 0.1 L < X_D < L, le décalage est faible. En régime d'inversion forte (voir Fig. (IV.7.a)), on constate la disparition de la double pente (due au changement de V_{FB}) lorsque la longueur de la zone chargée augmente.

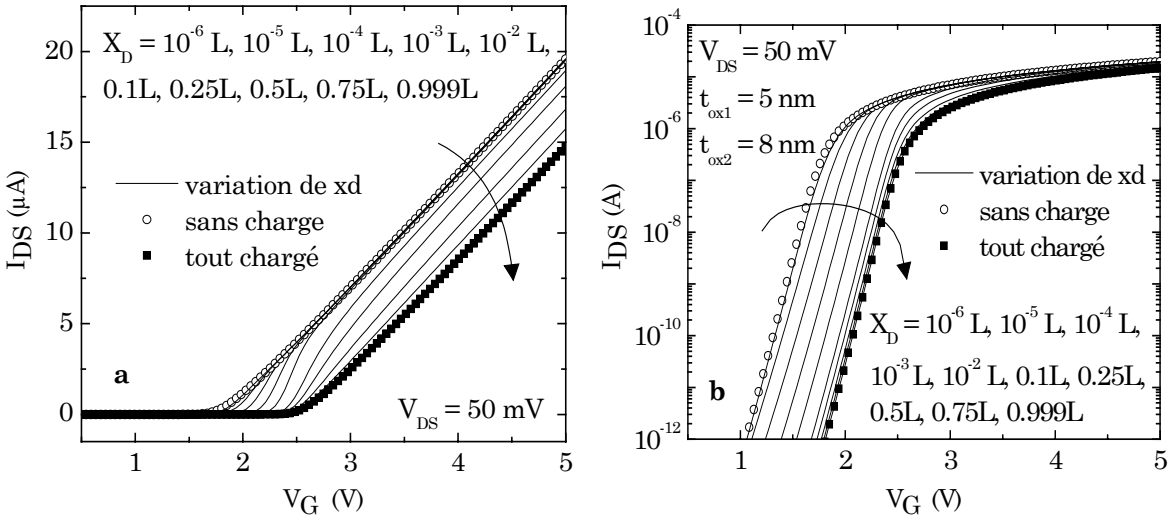

Figure IV.7. *Simulation de la courbe $I_{DS}(V_{GS})$ en fonction de la distance X_D proche du drain où sont localisées les charges fixes, en échelle linéaire (***a***) et en échelle logarithmique (***b***). Les paramètres de la simulation sont : W = 1 µm, L = 1 µm, t_{ox1} = 5 nm, t_{ox2} = 8 nm, $N_Q = 2\times10^{16}$ C.m^{-2} et V_{DS} = 50 mV.*

IV.4.2. Etude de la phase d'écriture des mémoires Flash à nodules

IV.4.2.1. Variation de la charge stockée

Dans un souci de simplification, nous montrons ici les résultats obtenus après découpage du transistor en deux transistors élémentaires. Même si pour une tension de grille donnée, le point de pincement se décale le long du canal en fonction de la polarisation appliquée sur le drain, nous supposons que l'injection du courant par porteurs chauds ne s'effectue que dans le second transistor (de longueur $X_D = 0.15 \times L$).

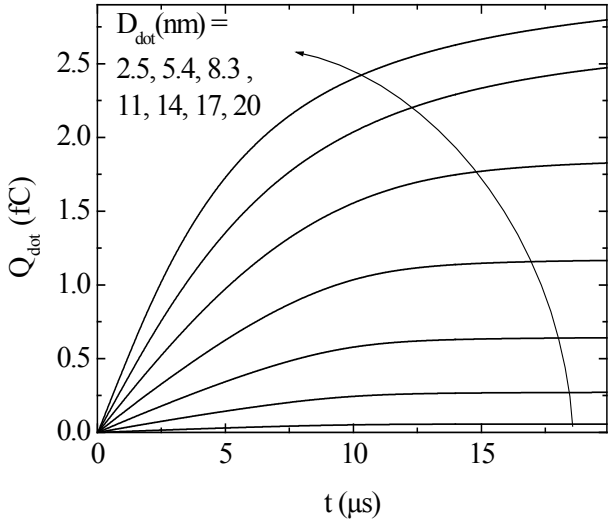

Figure IV.8. *Variation de la charge stockée, Q_{dot}, dans les nodules pendant l'écriture de la mémoire en fonction du diamètre des nodules. Les paramètres de la simulation sont : $W = 1\ \mu m$, $L = 1\ \mu m$, $t_{ox1} = 3$ nm, $t_{ox2} = 5$ nm, $V_{GB} = 5$ V, $V_{DS} = 3.5$ V, $X_D = 0.15 \times L$, $N_{dot} = 2 \times 10^{15}\ m^{-2}$ [Bernardini'03c].*

Nos simulations ont montré que le nombre d'électrons injectés par dot dépend des dimensions de ces derniers et de la durée de l'opération d'écriture. Un exemple de simulation dynamique des charges piégées dans les nodules (Q_{dot}) durant une injection CHE de 20 µs pour des nodules de différents diamètres est présenté sur la figure (IV.8). Cette figure met en évidence le lien direct entre la taille des nodules, le nombre d'électrons stockés et la durée de l'injection. Pour un temps d'écriture donné, l'augmentation de la taille des nodules induit une augmentation du nombre d'électrons stockés. Avec les paramètres choisis pour nos simulations, en imposant des conditions de polarisation, des temps d'écriture identiques et une densité de nodules faible ($2 \times 10^{11}\ cm^{-2}$), le nombre d'électrons stockés à la fin du temps d'écriture est respectivement de 1.2 et 58.3 électrons pour des diamètres de nodules de 2.5 nm et 20 nm. Par conséquent, pour conserver la même charge stockée lorsque le diamètre du nodule diminue, leur densité doit être augmentée.

IV.4.2.2. Décalage de la tension de seuil

Pour une tension de drain donnée, la détermination du décalage de la tension de seuil est réalisée à courant constant I_{test} (la valeur est choisie dans la pente sous le seuil) à l'aide d'un algorithme de calcul décrit sur la figure (IV.9). Pour une caractéristique $I_{DS}(V_{GS})$ donnée, cet algorithme permet de trouver les points pour lesquels les courants sont respectivement immédiatement supérieur et inférieur à la valeur I_{test}. Celle-ci étant choisie pour le régime de diffusion (pente sous le seuil), la droite qui joint les deux points encadrant la valeur de I_{test} est de la forme :

$$\text{Log } I_{DS} = a\, V_{GS} + b \qquad \text{(IV.8)}$$

La pente a et l'ordonnée à l'origine b de cette droite sont déterminées à partir des coordonnées des points trouvés autour de I_{test}. Par suite, la valeur de V_{GS} correspondant à I_{test} est obtenue, ainsi que le décalage de tension de seuil, pour des tensions de drain fixées ($\Delta V_{th} \cong \Delta V_{GS}$).

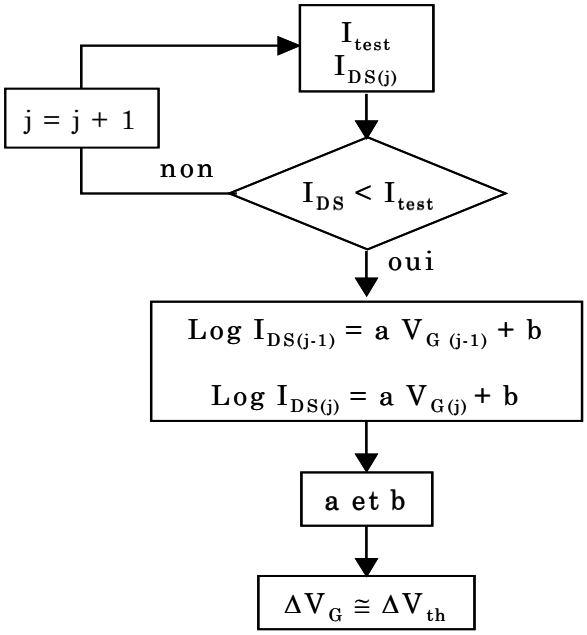

Figure IV.9. *Schématisation de l'algorithme de calcul pour la détermination du décalage de la tension de seuil.*

Les simulations $I_{DS}(V_{GS})$ effectuées pour différentes tensions V_{DS} présentées à la figure (IV.10), mettent en évidence l'influence de la tension appliquée sur le drain par rapport à l'amplitude du décalage de la tension de seuil, ΔV_T (déterminée à partir du décalage de la pente sous le seuil pour un courant I_{DS} fixé), pour différentes polarisations de la grille et différents diamètres de dots (D_{dot}). Cette figure (IV.10) montre également que pour une densité de dots fixée, N_{dot}, une petite réduction de leur diamètre entraîne une importante diminution de la tension de seuil d'écriture puisqu'il y a moins d'électrons piégés par nano-cristaux. De plus, la charge piégée augmente avec la polarisation de la grille (tant que le MOSFET est en régime de saturation) puisque la

quantité d'électrons injectés est plus importante. Le mécanisme de chargement s'arrête lorsque le champ électrique dans l'oxyde tunnel devient négligeable : $\Delta V_T(V_{DS})$ tend à saturer pour des valeurs de V_{DS} plus élevées.

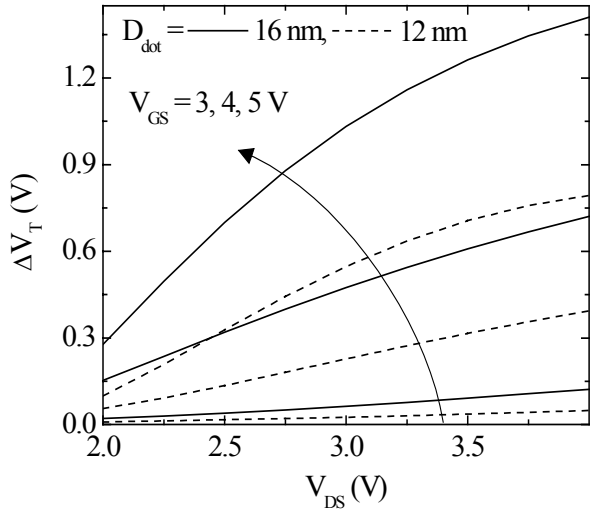

Figure IV.10. *Simulation de l'évolution de la tension de seuil en fonction de la tension appliquée sur le drain pour différentes tensions de grille et des dots de différents diamètres. Les paramètres du transistor sont les mêmes que ceux reportés sur la figure (IV.8) avec une période d'écriture de 10µs [Bernardini'03c].*

IV.5. Caractérisations électriques de structures avec nodules

Récemment, de nouveaux procédés de fabrication des mémoires à nodules ont permis une amélioration du contrôle de la densité (N_{dot}) et du diamètre (D_{dot}) des nodules de silicium [De Salvo'03]. Dans la suite de ce chapitre, nous présenterons l'étude menée sur des dispositifs décrits dans cette publication, à savoir, quatre demi-plaques de silicium ayant toutes des nodules de silicium de tailles et de densités différentes permettant de garder cependant un coefficient R_{eff} constant d'environ 25% (voir tableau (IV.1)).

Demi-plaque	N_{dot} (10^{15} m^{-2})	D_{dot} (nm)	R_{eff} calculé (%)
1	16	4.5	25.4
2	9.6	5.5	22.8
3	4	8.5	22.7
4	2.8	10	22

Tableau IV. 1. *Récapitulatif des caractéristiques des 4 demi-plaques fabriquées par ST Microelectronics Catagne [De Salvo'03].*

Le manque de plaque témoin (sans nodule) de ce lot, nous a orienté vers des études comparatives entre plaques.

IV.5.1. Etude des capacités avec nodules

Dans un premier temps, nous nous sommes intéressés aux capacités avec nodules pour extraire ou vérifier les paramètres caractéristiques de ces structures.

IV.5.1.1. Etude expérimentale

Différentes mesures capacitives quasi-statiques ont été réalisées sur des capacités à nodules de différentes surfaces, présentes sur les quatre demi-plaques dont nous disposions. La figure (IV.11) présente les mesures quasi-statiques obtenues pour des capacités, de surface A_{eff} = 3 10^{-3} cm², situées dans trois régions différentes des plaques (au milieu, à droite et à gauche). Le faible décalage entre les valeurs maximales des capacités mesurées au milieu, à droite et à gauche des demi-plaques, témoigne de la bonne uniformité des plaques. Pour les tensions de grille positives, on observe un début de chute de la valeur de la capacité puis la courbe tend à saturer.

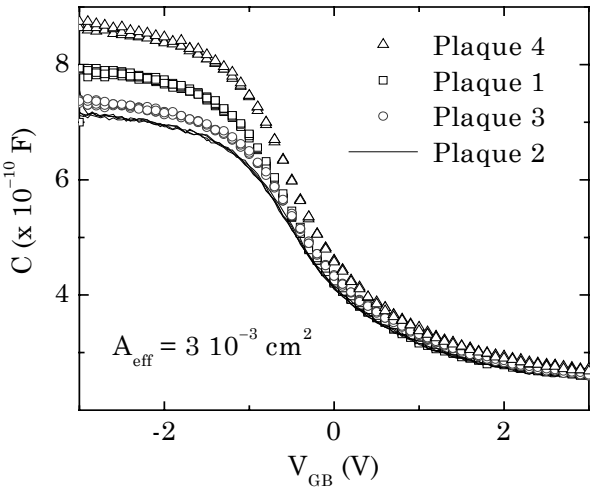

Figure IV.11. *Mesures C-V quasi-statiques effectuées sur les capacités à nodules de surface A_{eff} = 3 10^{-3} cm².*

On peut également remarquer que l'ordre des courbes tracées ne correspond pas à la réduction de R_{eff} donnée dans le tableau (IV.1), puisque les courbes correspondant à la plaque 1 sont situées entre celles de la plaque 3 et de la plaque 4.

IV.5.1.2. Modélisation des capacités à nodules

Afin d'expliquer l'ordre des courbes C-V mesurées, nous avons modifié les programmes développés lors de l'étude de la capacité MOS pour simuler des capacités à nodules (C_{dot}) en tenant compte du facteur R_{eff} :

$$C_{dot} = \frac{A_{eff}}{\frac{1}{C_{SC}} + \frac{t_{ox1}}{\varepsilon_{ox}\varepsilon_0} + \frac{t_{ox2}}{\varepsilon_{ox}\varepsilon_0} + \frac{\frac{D_{dot}}{2}}{\varepsilon_{Si}\varepsilon_0 R_{eff}}} \qquad \text{(IV.9)}$$

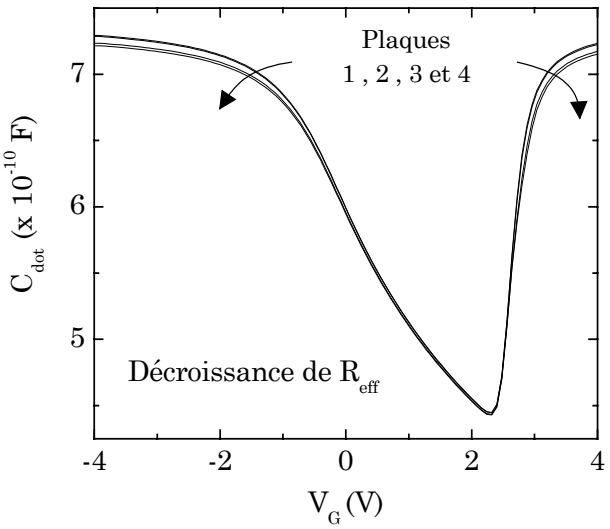

Figure IV.12. *Simulations des courbes C-V pour les plaques 1, 2 3 et 4 avec les coefficients R_{eff} donnés dans le tableau (IV.1). Les paramètres utilisés pour les simulations sont : N_A = 1.3 10^{24} m^{-3}, t_{ox1} = 5.5 nm, t_{ox2} = 8 nm, A_{eff} = 3 10^{-3} cm^2 et V_{FB} = –1 V.*

Contrairement aux courbes C-V mesurées (cf. Fig. (IV.11)), les simulations réalisées pour les valeurs de R_{eff} reportées dans le tableau (IV.1), montrent une décroissance continue des valeurs maximales des capacités (voir Fig. (IV.12)). Par conséquent, les coefficients R_{eff} calculés à partir du diamètre et de la densité des nodules ne permettent pas d'expliquer l'ordre des courbes de la figure (IV.11). Nous nous sommes donc intéressés plus particulièrement à l'impact du diamètre des nodules.

IV.5.1.3. Impact de la densité, N_{dot}, et du diamètre, D_{dot}, des nodules

Dans un premier temps, nous avons simulé les courbes C-V des capacités avec et sans couche de semiconducteur, à l'intérieur de la couche d'oxyde (c.a.d. deux ou trois capacités en série). La figure (V.13) met en évidence la diminution de la capacité lorsque l'épaisseur de la couche de silicium augmente à l'intérieur de l'oxyde.

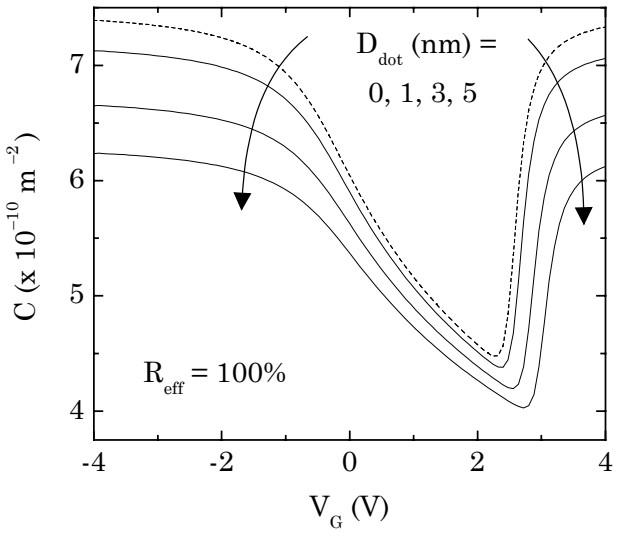

Figure IV.13. *Simulation des courbes C-V sans ou avec une couche de silicium à l'intérieur de la couche d'oxyde pour R_{eff} = 100%. Les paramètres de la simulation sont : A_{eff} = 3 10^{-7} m^{-2}, t_{ox1} = 5.5 nm, t_{ox2} = 8 nm, N_A = 1.3 10^{24} m^{-3}, N_{dot} = 10^{11} cm^{-2} et V_{FB} = –1 V.*

Cependant, les nodules ne recouvrent qu'une partie R_{eff} de la surface de l'oxyde. Par conséquent, les courbes C-V correspondant aux capacités avec nodules (C_{dot}), se situent entre les courbes C-V simulées pour $D_{dot} = 0$ nm et $D_{dot} \neq 0$ nm. Néanmoins, à cause des différentes valeurs de R_{eff} (dues à N_{dot} et D_{dot}), les courbes C-V de capacités ayant des nodules de diamètres différents pourront se superposer ou ne pas suivre l'ordre croissant de la taille des nodules. En d'autres termes, les valeurs de N_{dot} et D_{dot} données dans le tableau (IV.1) ne correspondent pas de façon assez précise aux dispositifs mesurés.

De plus, on constate que les valeurs des capacités mesurées (cf. Fig. (IV.11)) sont plus faibles que celles obtenues par simulations (cf. Fig. (IV.13)). Par conséquent, l'épaisseur d'oxyde utilisée pour les simulations ne correspond pas à celle des structures mesurées.

La figure (IV.14) met en évidence le décalage des courbes C-V des capacités à nodules pour différentes valeurs d'épaisseur d'oxyde (t_{ox1} et t_{ox2}). La comparaison entre les mesures, figure (IV.11), et les simulations, figure (IV.14), montrent que lors de nos premières simulations, nous avons surestimé les valeurs des épaisseurs d'oxyde t_{ox1} et /ou t_{ox2}. Cette imprécision sur les épaisseurs peut également expliquer l'ordre des courbes mesurées en considérant que les quatre demi-plaques n'ont pas exactement la même épaisseur d'oxyde (tout en supposant qu'elles aient le même dopage de substrat).

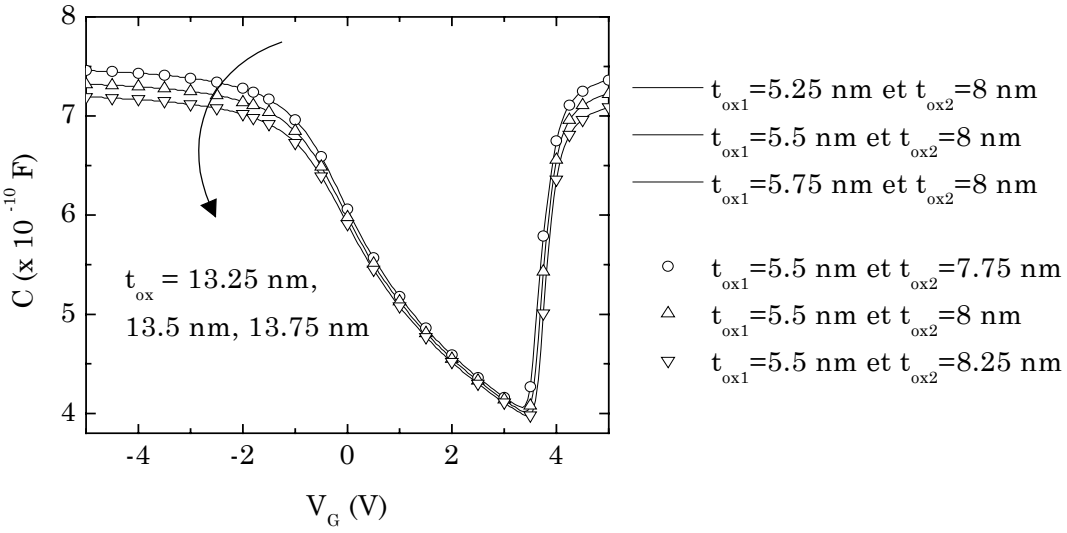

Figure IV.14. *Simulation des courbes C-V de capacités à nodules en fonction de la variation des épaisseurs d'oxydes t_{ox1} et t_{ox2}. Les paramètres de simulations sont : $N_A = 1.3\ 10^{24}\ m^{-3}$, $A_{eff} = 3\ 10^{-7}\ m^2$, $V_{FB} = -1$ V, $N_{dot} = 16\ 10^{15}\ m^{-2}$, $D_{dot} = 4.5$ nm.*

IV.5.1.4. Extraction des paramètres des capacités ring

Les programmes d'extraction de paramètres (dopage de substrat, épaisseur d'oxyde, tension de bandes plates) développés pour nos travaux de recherche sur la capacité MOS

(cf. chapitre II) ont été adaptés au cas des capacités à nodules en tenant compte de l'équation (IV.8) pour simuler la capacité avec dots. La figure (IV.15) donne un exemple de l'extraction des paramètres d'une capacité ring, de surface A_{eff} = 5 10^{-3} cm², située sur la demi-plaque 1 décrite dans le tableau (IV.1). Pour des tensions comprises entre la tension de bandes plates et la tension de seuil, on observe sur la figure (IV.15.a) un croisement de la courbe mesurée et de celle simulée, qui met en évidence la non uniformité du dopage de substrat.

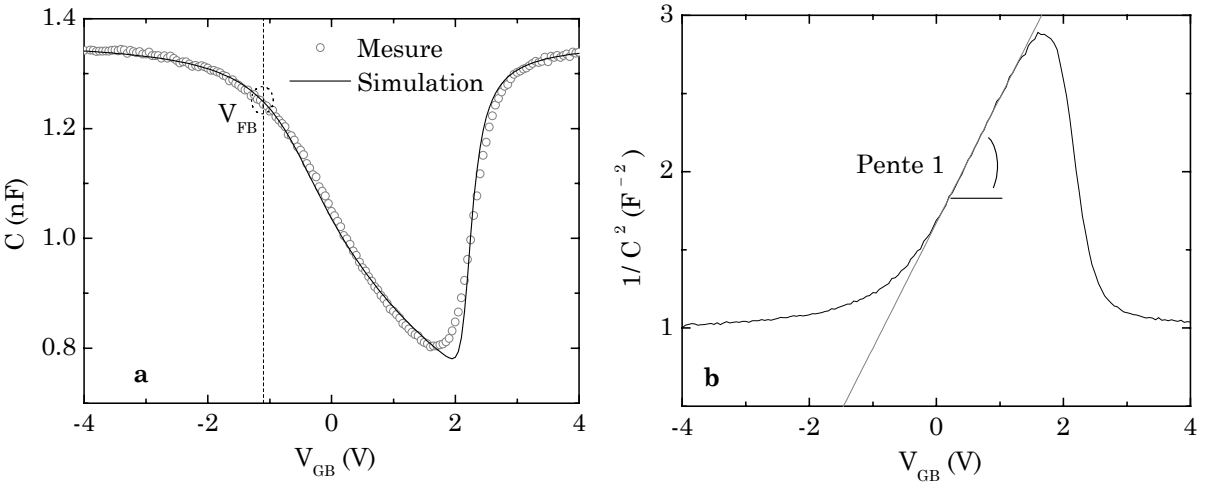

Figure IV.15. *Comparaison entre la courbe C-V mesurée et simulée pour une capacité à nodules de la demi-plaque 1 (A_{eff} = 5 10^{-3} cm²)* (**a**) *et extraction de ses paramètres par exemple Na (***b**). *Les différents paramètres extraits sont : N_A = 1.29 10^{24} m^{-3}, t_{ox} = 12.4 nm, V_{FB} = – 1.1 V.*

Le tableau (IV.2) résume les valeurs des paramètres extraits à partir des mesures C-V des capacités ring, de surface A_{eff} = 5 10^{-3} cm², situées sur les quatre demi-plaques étudiées.

Demi-plaques	N_A (10^{24} m^{-3})	t_{ox} (nm)	V_{FB} (V)
1	1.29	12.4	– 1.1
2	1.32	13.76	
3	1.33	13.32	
4	1.27	11.3	

Tableau IV. 2. *Récapitulatif des paramètres extraits à partir des courbes C-V des capacités ring de surface Aeff = 5 10^{-3} cm², pour les quatre demi-plaques étudiées.*

Ces résultats mettent en évidence la variation des paramètres clefs des capacités entre les quatre demi-plaques.

IV.5.2. Etude des transistors avec nodules

IV.5.2.1. Etude des temps d'écriture

L'écriture par porteurs chauds permet la diminution du temps d'écriture des mémoires. De plus, les simulations, présentées sur la figure (IV.8), ont mis en évidence la saturation de la charge injectée au bout de quelques dizaines de micro-secondes suivant le diamètre des nodules. A partir de ces observations, nous avons procédé à l'opération d'écriture d'un même dispositif en polarisant la grille à 8 V et le drain à 3.5 V avec un ou plusieurs pulses de 500 µs, 50 µs et 5 µs. Avant chaque écriture, nous avons pris soin d'effacer la cellule par injection FN en appliquant une tension nulle sur le drain, la source et le bulk et en appliquant une tension de – 12 V sur la grille pendant 100 ms.

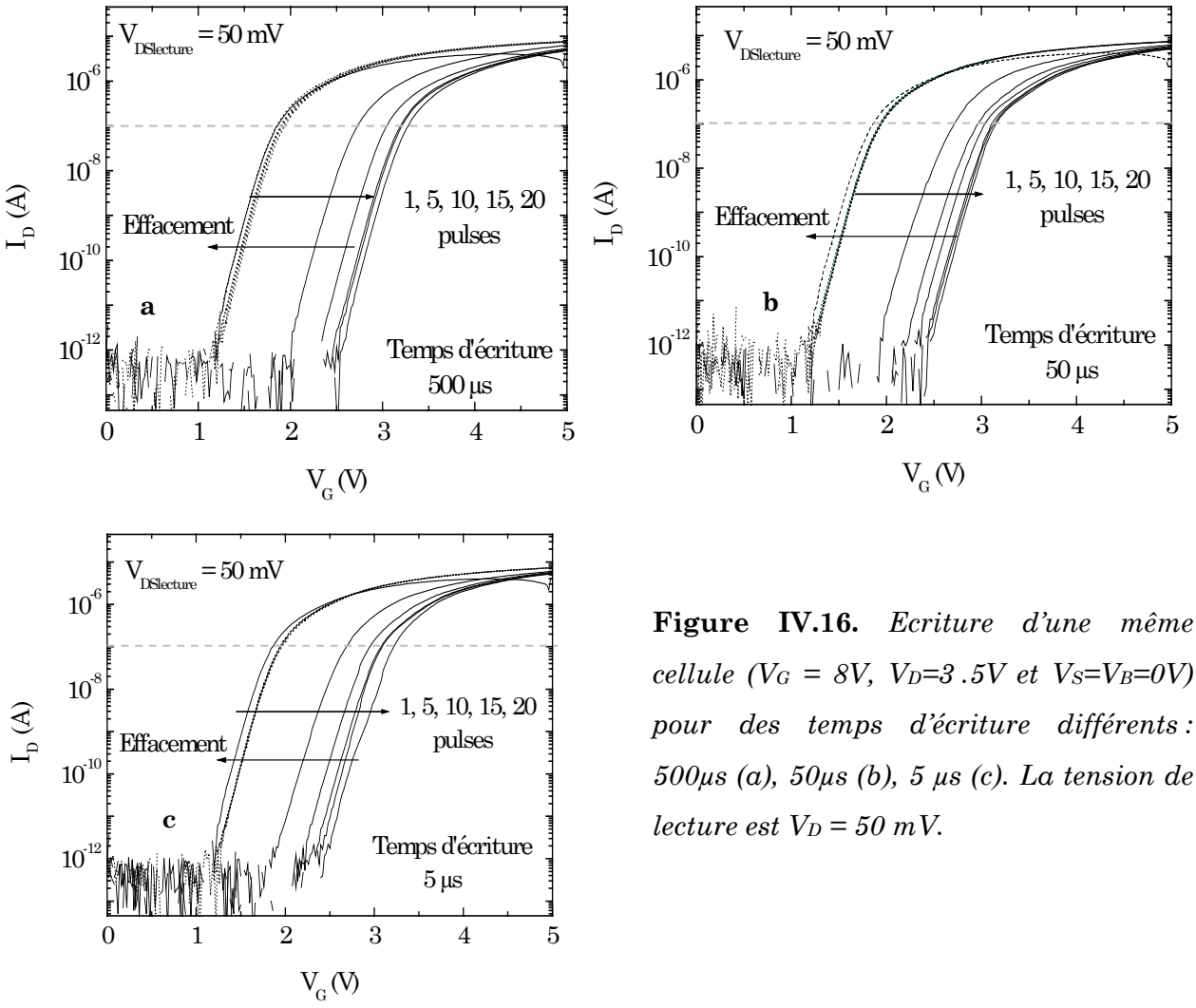

Figure IV.16. *Ecriture d'une même cellule (V_G = 8V, V_D=3 .5V et V_S=V_B=0V) pour des temps d'écriture différents : 500µs (a), 50µs (b), 5 µs (c). La tension de lecture est V_D = 50 mV.*

La figure (IV.16) montre que les courbes correspondant à l'état effacé sont quasiment confondues (un léger décalage peut apparaître suite à la génération d'états d'interface lors de la première écriture de la mémoire) alors qu'après les opérations d'écriture les courbes se décalent de façon similaire. L'extraction des tensions de seuil à courant fixé (I_{test} = 10^{-7}A) présentée à la figure (IV.17), met en évidence le faible décalage entre les tensions de seuil obtenues pour des temps d'écriture de 500 µs, 50 µs et 5 µs, ce qui témoigne du temps très court nécessaire à l'injection d'électrons dans les nodules.

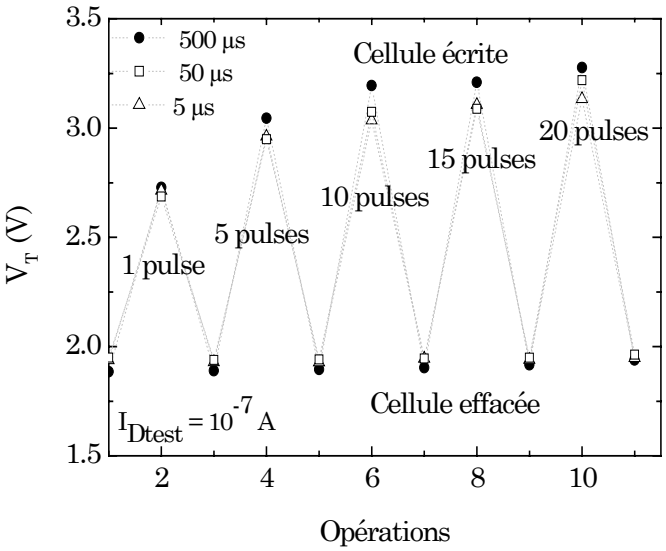

Figure IV.17. *Evolution de la tension de seuil des courbes tracées sur la figure (IV.15).*

IV.5.2.2. Etude des tensions d'écriture

Pour une tension de grille donnée, l'injection par porteurs chauds se produit pour une tension de drain environ égale à la moitié de cette tension de grille. Cependant dans le but de trouver les conditions d'écriture optimales, à savoir un compromis entre la génération de paires électron-trous et une faible dégradation de l'oxyde, différents couples (V_{GS}, V_{DS}) ont été étudiés. Par exemple, la figure (IV.18) montre que pour une tension de grille d'écriture égale à 8 V, la tension de seuil commence à se décaler à partir d'une tension de drain de lecture de 2 V, puis le décalage est moins prononcé à partir de 5 V avant de revenir en arrière pour des tensions supérieures à 6 V. Ce phénomène peut être dû à la dégradation de l'oxyde ; en effet, lors de ces expériences, nous avons observé le claquage des échantillons testés dès que la tension de seuil extraite commençait à décroître (entre 6 V et 7 V). Afin de vérifier et valider le comportement de notre simulateur lors de l'opération d'écriture, nous avons mesuré les caractéristiques courant-tension des mémoires à nodules après différentes programmations.

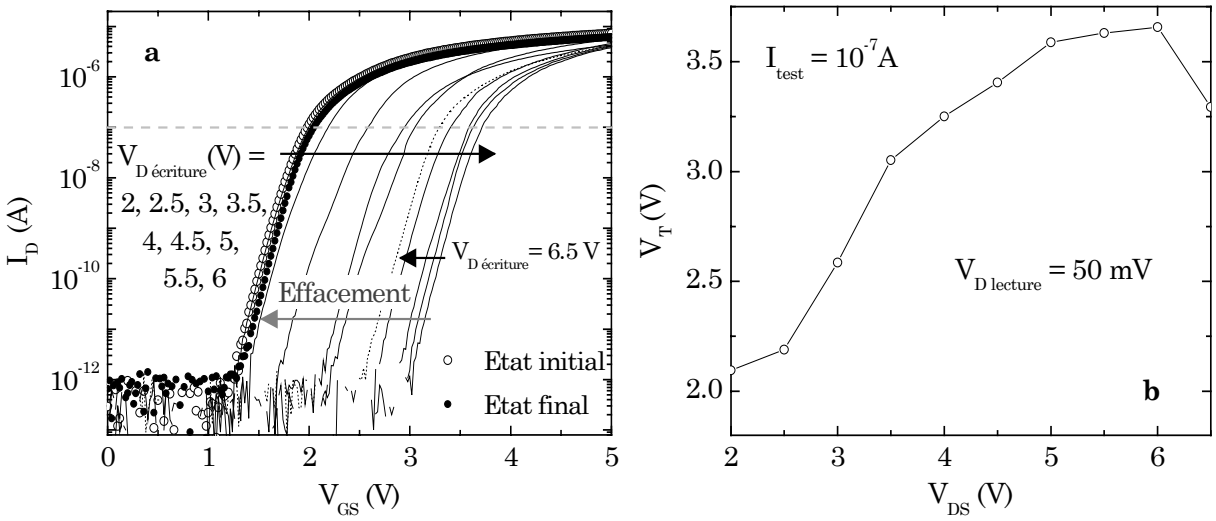

Figure IV.18. *Lecture d'un transistor à nodules pour une tension V_D = 50 mV après différentes écritures cumulatives avec une tension de grille fixée à 8 V et une tension de drain croissant de 2 V à 6.5 V durant 50 µs (***a***) et extraction de la tension de seuil correspondante, à courant fixé ($I_{test} = 10^{-7}$ A) (***b***).*

La figure (IV.19) présente la variation de la tension de seuil extraite, à courant fixé ($I_{test} = 10^{-7}$ A), à partir des mesures $I_D(V_{GS})$ effectuées après l'écriture de la mémoire à différentes polarisations de grille et de drain. On observe le même comportement électrique que celui obtenu avec notre simulateur à savoir un grand décalage de la tension de seuil induit par l'augmentation des tensions de grille et de drain (voir Fig. (IV.10)).

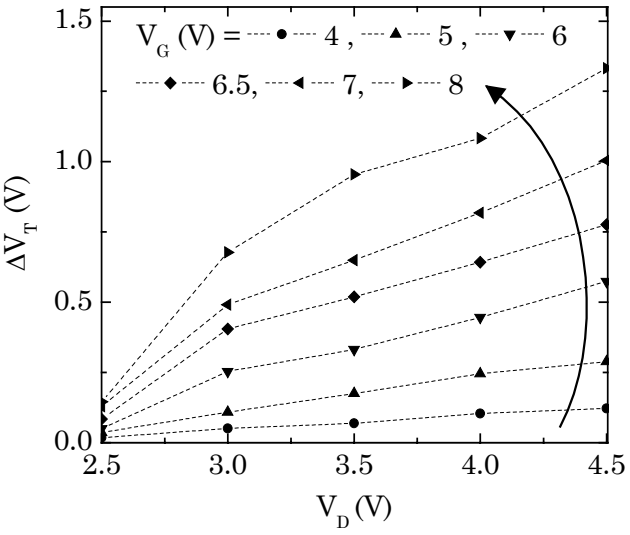

Figure IV.19. *Variation de la tension de seuil d'un transistor (W = 0.16 µm, L = 0.28 µm) situé sur la demi-plaque 2, en fonction du potentiel de drain pour différentes tensions de grille appliquées durant un temps d'écriture égal à 10 µs (la lecture se fait à V_D = 50 mV) .*

.

Pour cette gamme de tension, la mémoire reste en régime de saturation et l'injection du courant augmente avec la tension appliquée sur la grille. Pour des temps d'écriture plus longs, la courbe ΔV_T (V_{DS}) aurait tendance à atteindre une valeur constante, c.a.d. une valeur de saturation. Bien qu'il ne nous ait pas été possible de calibrer notre

simulateur sur les dispositifs (en particulier pour les paramètres CHE), par manque de transistors de test (sans nodule), on peut noter, au premier ordre, une bonne concordance entre simulations et mesures ce qui tend à valider notre approche.

IV.5.2.2. Etude de la tension de lecture

Afin de ne pas écrire pendant l'opération de lecture, il est nécessaire de connaître la tension de drain à partir de laquelle l'injection CHE se produit. Différentes séries de mesure $I_D(V_{GS})$ ont donc été effectuées en augmentant la tension de drain de 50 mV à 4 V, tout en gardant les tensions de substrat et de bulk nulles. La figure (IV.20) présente les différentes variations des courants en fonction de la polarisation de grille lorsque la tension de drain V_D augmente.

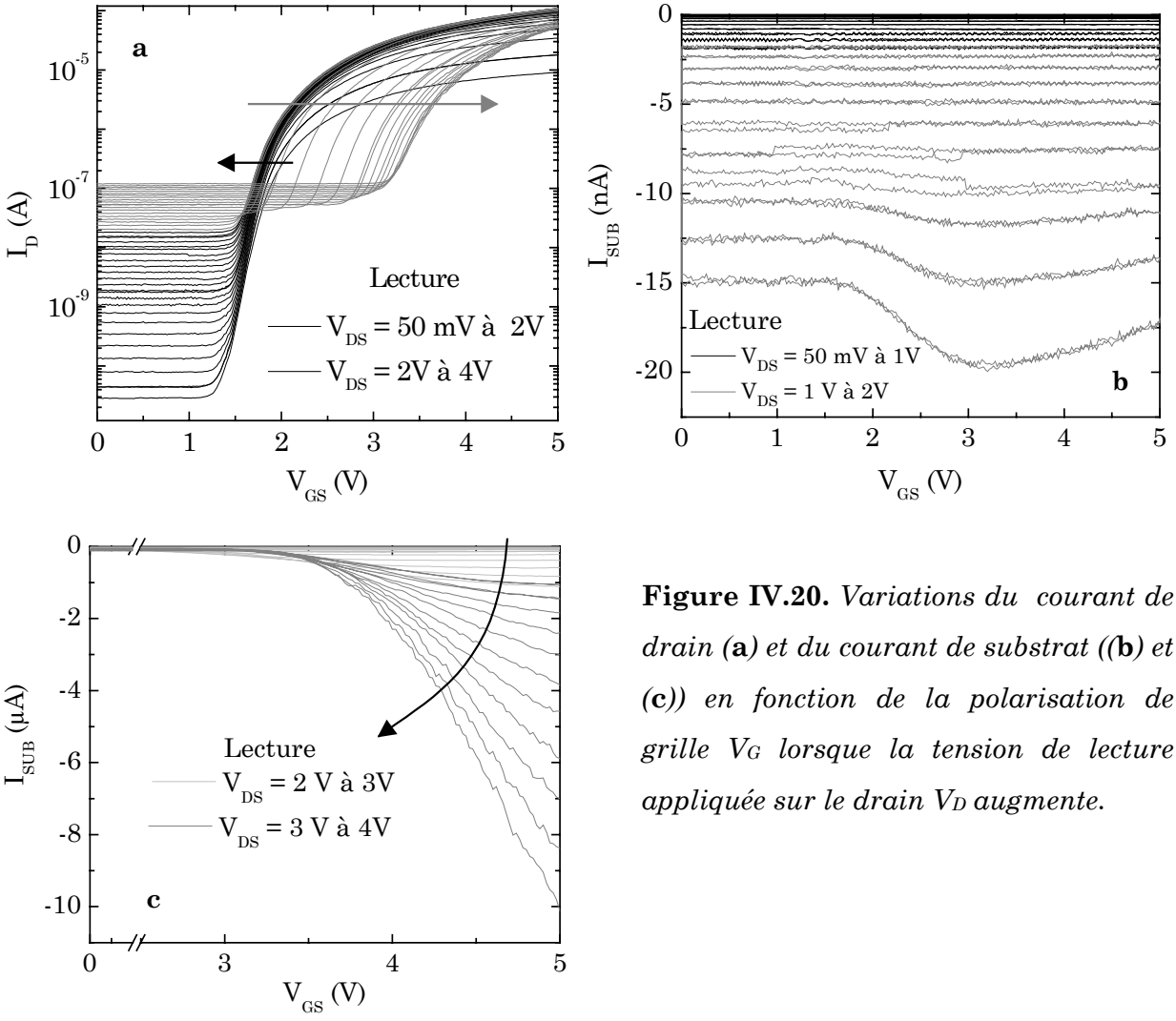

Figure IV.20. *Variations du courant de drain (***a***) et du courant de substrat ((***b***) et (***c***)) en fonction de la polarisation de grille V_G lorsque la tension de lecture appliquée sur le drain V_D augmente.*

On observe un décalage des courbes I_D (V_{GS}) du côté des tensions de grille positives dès que V_D devient supérieur à 2 V, ce qui correspond sur la figure (IV.20.b) au début de l'augmentation du courant de substrat, c'est à dire au début de l'injection par porteurs

chauds. De surcroît, la figure (IV.20.c) met en évidence l'augmentation continue du courant de substrat pour les fortes tensions de drain. Jusqu'à présent, seule la tension de lecture appliquée sur le drain a été présentée. Toutefois, les dispositifs dont nous disposons, présentent un coefficient R_{eff} faible de l'ordre de 25%, ce qui leur confère la possibilité d'un fonctionnement 2 bits. En effet, Mulidhar *et al.* ont montré que tant que le coefficient R_{eff} restait inférieur à la valeur critique du seuil de percolation, les îlots de silicium étaient suffisamment isolés les uns des autres pour éviter le transport latéral des charges entre nodules [Muralidhar'03]. Pour observer ce fonctionnement 2 bits, et vérifier que l'injection est bien localisée d'un seul côté, il est nécessaire de lire la mémoire à la fois en mode direct ($V_D > 0$ et $V_S = 0$) pour déterminer la tension de seuil côté source (V_{thf}) et en mode inverse (c.a.d. en inversant les polarisations source et drain ($V_D=0$ et $V_S>0$)), pour déterminer la tension de seuil du côté drain (V_{thr}). Cependant, les lectures effectuées pour des tensions égales à 50 mV ne permettent pas de faire la distinction entre les deux tensions de seuil. Ce phénomène, observé par Bloom *et al.* [Bloom'02] pour les mémoires à nitrure, est dû à la valeur de la tension de lecture. En effet, pour observer le fonctionnement deux bits des mémoires, il est nécessaire de placer le dispositif en régime de saturation.

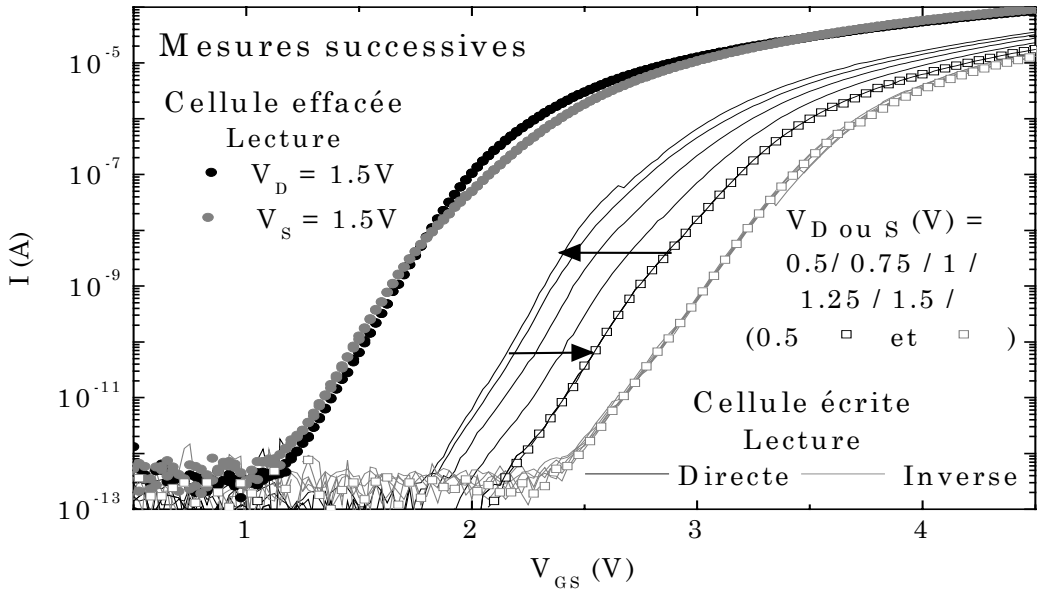

Figure IV.21. *Mesures successives des caractéristiques courant-tension d'une cellule (de la demi-plaque 2). Dans un premier temps la cellule a été effacée puis lue en mode direct (points noirs) et en mode inverse (points gris) pour des tensions de lecture de 1.5V ; puis dans un deuxième temps la cellule a été programmée (V_G = 8Vet V_D = 3.5V durant 500µs) et lue en mode direct (traits noirs) et en mode inverse (trait gris) pour des tensions de lectures variant de 0.5V à 1.5V et à nouveau 0.5V.*

Comme le montre la figure (IV.21), à la suite de lectures successives, une fois la cellule écrite (écriture V_G = 8V, V_{DS} = 3.5V pendant 500µs), plus la tension de drain augmente, plus les caractéristiques $I_D(V_{GS})$ se décalent vers la caractéristique $I_D(V_{GS})$ de la cellule effacée. Notons, que la cellule mesurée n'a pas été dégradée par la série de lectures puisque les deux caractéristiques $I_D(V_{GS})$ lues en mode direct et inverse pour $V_{lecture}$ = 0.5V avant et après cette étude sont superposables. Le décalage des courbes $I_D(V_{GS})$ dû à la tension de lecture V_D, résulte de la présence des effets canaux courts, et plus particulièrement de l'effet DIBL (Drain Induced Barrier Lowering). En effet, la longueur de la zone d'injection étant très courte [Shappir'03], lorsque la tension appliquée sur le drain augmente, la couche de déplétion s'étend de plus en plus dans le canal vers la source et il se produit un abaissement de la barrière source/canal. L'abaissement de la barrière à la source permet l'injection d'électrons au travers du canal (en surface) et ceci indépendamment de la tension de grille. Par suite, en régime sous seuil, la grille perd le contrôle du courant de drain.

De surcroît, la tension de seuil obtenue par une lecture en mode direct est plus faible que la tension de seuil obtenue par une lecture en mode inverse [Eitan'00], [Bloom'01] et [Larcher'02]. Si la charge est injectée côté drain, elle sera complètement écrantée par la forte valeur appliquée en mode de lecture direct. Par suite, la forte polarisation V_D induit une région de pincement au dessous de la zone chargée (zone où il n'y a pas de couche d'inversion) et la caractéristique $I_D(V_{GS})$ de la cellule écrite reste proche de la caractéristique de la cellule vierge. Par conséquent la tension de seuil reste faible. En revanche, durant le mode de lecture inverse, la forte tension appliquée sur la source n'est pas capable d'écranter l'effet des électrons. Ainsi, pour étudier le fonctionnement 2 bits des cellules mémoires à nodules, des tensions de lecture supérieures à 1V doivent être appliquées tout en restant au dessous de la tension où débute l'injection par porteurs chauds (environs 2V).

Enfin, la dégradation de la pente sous le seuil, surtout visible en mode de lecture inverse, prouve également la présence des effets 2D qui apparaissent à côté de la jonction où s'effectue l'injection. Cette dégradation est due à la courte longueur de la zone chargée par injection CHE plutôt qu'à la génération d'états d'interfaces durant l'injection [Shappir'03].

IV.6. Conclusion

Le travail présenté dans ce chapitre, repose sur les modèles préalablement développés lors de nos travaux de recherches sur les structures MOS. En modifiant ces modèles pour les adapter au cas particulier des mémoires à nodules, nous avons mis en évidence l'impact d'une charge piégée non uniformément dans tous les régimes (faible à forte inversion) de fonctionnement. Puis nous avons développé un algorithme de calcul permettant la modélisation électrique statique et dynamique de ces mémoires. Ce modèle peut également être une aide au design ou pour l'optimisation des signaux d'écriture et d'effacement des mémoires.

Les différentes comparaisons entre les mesures et les simulations des capacités à nodules témoignent de la complexité de la modélisation de ces structures due aux petites variations des valeurs des paramètres clefs (épaisseur d'oxyde, dopage, diamètres des nodules, densité des nodules).

Enfin, bien qu'il ne nous ait pas été possible de calibrer notre simulateur sur les dispositifs par manque de transistors de test (sans nodule), les caractérisations des transistors à nodules ont permis de montré un bon comportement de notre modèle pour une étude statique 1bit des mémoires à nodules. Cependant, les caractérisations électriques des mémoires à nodules lors d'une étude 2 bits montrent une dégradation de la pente sous le seuil qui selon Lusky *et al.* [Lusky'01] - [Lusky'04] est due aux charges piégées au dessus du drain et du canal proche du drain et aux variations des effets 2D du champ électrique dans cette zone. Ainsi notre modèle, qui ne prend pas en compte les phénomènes 2D tels que les effets canaux courts, n'est plus valide pour ce mode de fonctionnement.