Développement de mémoires non-volatiles embarquées pour les plateformes technologiques avancées 40nm et 28nm

M. Hesse[a,b], A. Regnier[a], P. Masson[b]

[a] STMicroelectronics Rousset, 190 avenue Célestin Coq, 13106 Rousset, France
[b] Laboratoire EpOC – URE UNS 006, Ecole Polytechnique de l'Université Nice Sophia Antipolis, 930 route des Colles – Campus Sophiatech, 06410 Biot, France
Contact email : marjorie.hesse@st.com

Ce papier décrit le développement des prochaines générations de mémoires non volatiles (NVM). Ce travail se focalise dans un premier temps sur un rappel des technologies mémoires embarquées dites classiques, intervenant sur de nombreuses applications à base de microcontrôleurs. L'accroissement de ce champ d'applications engendre le respect de nombreux critères (ultra basse consommation, augmentation de puissance, etc.) nécessitant une évolution technologique vers des nœuds avancés comme le 40nm et le 28nm. Un état de l'art des architectures actuelles permettra alors de découvrir les différentes technologies existantes et de comprendre les procédés utilisés. Nous aborderons d'abord les notions théoriques utilisées, puis les freins potentiels à anticiper dans cette génération de mémoire et ses possibles améliorations.

I. Introduction

Le développement massif des appareils électroniques nomades (téléphone portable, tablette, montre connectée, implant biomédical, etc.) accroit le champ d'applications des microcontrôleurs montré sur la Figure 1.

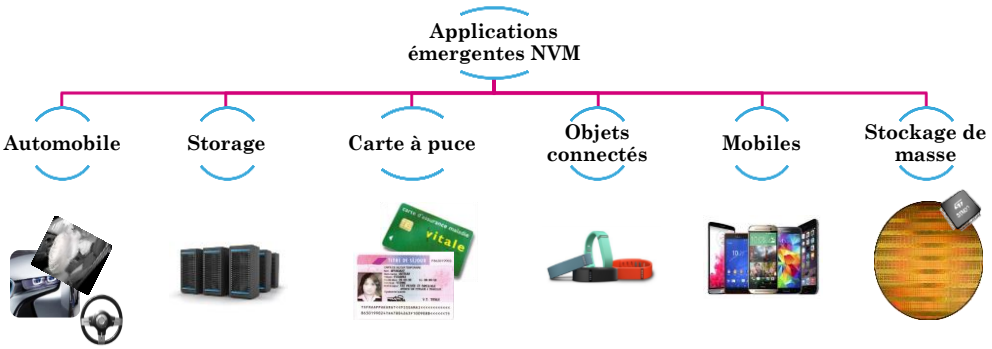


Fig.1. Applications émergentes des mémoires non volatiles

Ceci s'accompagne d'une augmentation de la puissance consommée limitant l'autonomie, d'une plus grande vitesse de traitement, du critère de densité et de plus fortes capacités de stockage en mémoire non volatile programmable. Deux types de mémoires sont alors distinguables : la mémoire autonome (stand-alone memory) et la mémoire embarquées (embedded memory). La mémoire autonome occupe de nos jours la majorité du marché et est utilisée dans une grande quantité de produits comme la clé USB ou le disque dur. En ce qui concerne les mémoires embarquées (eNVM), il en existe deux types dépendant du niveau d'intégration de la mémoire : le SiP (System in Package) et le SoC

Article publié par EDP Sciences et disponible sur le site https://www.j3ea.org ou https://doi.org/10.1051/j3ea/20171003

(System on Chip). Le SoC est la cible des NVM émergeantes suite aux récentes révolutions technologiques sollicitant en particulier les microcontrôleurs. C'est pourquoi l'eNVM connait un marché en expansion poussant son développement au-delà des limites.

II. L'évolution de l'eNVM

Les microcontrôleurs ultra basse consommation sont basés sur différents types de technologies d'eNVM dites classiques : Flash NOR (1), cellule Split Gate (2), SuperFlash (3,4), SONOS (5) ou encore EEPROM (6).

L'architecture Flash NOR classique offre une bonne intégration grâce à l'implémentation « 1 transistor par bit ». Nous retrouvons les technologies des mémoires Flash embarquées classiques en Figure 2. Son écriture en porteurs chauds (7) présente cependant une consommation en écriture assez élevée. De plus, la circuiterie périphérique d'écriture est complexe (algorithme d'écriture embarqué) et occupe une place considérable (capacités des pompes de charge). Cependant, son architecture et sa géométrie sont exploitables afin d'en faire une cellule de plus petite taille, ce que nous montrerons par la suite.

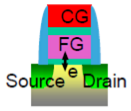
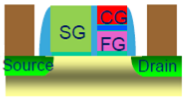
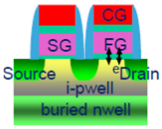
Type	1 Transistor	1.5 Transistors	2 Transistors
Structure			
Mode de programmation	Porteurs chauds (CHE)	Injection côté source (SSI)	Porteurs chauds (CHE)
Mode d'effacement	Fowler-Nordheim (FN)	Fowler-Nordheim (FN)	Fowler-Nordheim (FN)
Avantage	Petite taille	Programmation rapide	Basse consommation
Inconvénient	Haute consommation	Difficulté d'intégration	Grande taille

Fig.2. Technologies des mémoires Flash embarquées classiques.

Une autre approche plus adaptée aux basses densités est la mémoire à injection de type Fowler-Nordheim (8) comme l'EEPROM ou la Flash UCP (9). Il s'agit ici d'une architecture composée de deux transistors par bit, avec une périphérie simple et dense. Cependant, le point mémoire étant quatre à cinq fois plus gros que la Flash, elle n'est pas adaptée aux blocs de taille supérieure à 200 kilo-octets.

C'est pourquoi les concurrents choisissent de préférence une Flash à grille partagée, plus généralement nommée *split-gate*. Cette architecture est un compromis entre la Flash et l'EEPROM. Néanmoins, elle est relativement complexe et impose une fabrication délicate. De plus, elle ne permet pas de répondre à tous les critères exigibles pour les microcontrôleurs nécessitant une ultra-basse consommation car les tensions d'activation restent élevées, de l'ordre de 5V à 15V.

Pour les nœuds technologiques avancés comme le 40nm, la plupart des fournisseurs continuent à travailler sur ces architectures conventionnelles. Ces solutions techniques correspondent essentiellement à des architectures mémoire dans le FEOL (Front End Of Line) ou à stockage de charges (floating gate) permettant ainsi l'utilisation de matériaux conventionnels.

Afin de réduire l'énergie nécessaire à la programmation et à l'effacement de ces points mémoires, certains fournisseurs de microcontrôleurs développent de nouveaux types de mémoire qui s'intègrent au BEOL (Back End Of Line) (10). Cela concerne par exemple les mémoires magnéto résistives STT-MRAM, à changement de phase PCM, les mémoires

ferroélectriques FRAM ou encore les mémoires résistives RRAM, que nous pouvons retrouver sur la Figure 3. Cependant, ces mémoires alternatives ne répondent pas à tous les critères d'utilisation des microcontrôleurs ultra-basse consommation, bien souvent à cause de leur courant de programmation élevé. Ces concepts présentent aussi une forte rupture technologique, notamment avec l'intégration de nouveaux matériaux.

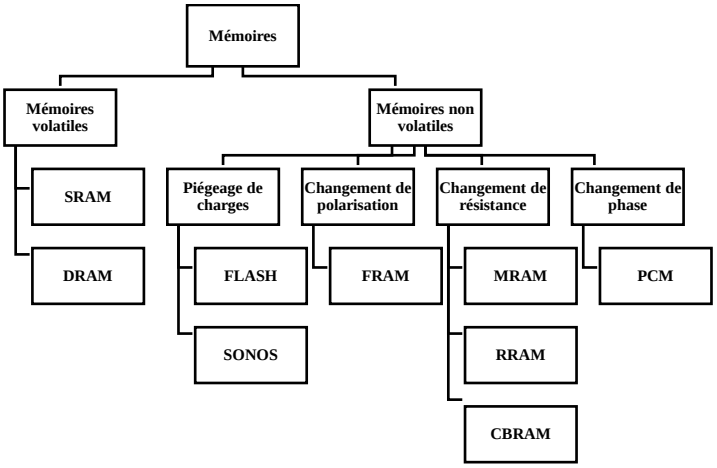


Fig.3. Classification succincte des mémoires

Pour le nœud technologique 28nm, la définition de la plateforme logique nécessite l'introduction de nouveaux matériaux comme le substrat SOI, les grilles métalliques et les oxydes à forte permittivité (high-k) tels que le HfSiON. Ces matériaux complexifient l'intégration de cellule NVM et cela sans impacter le fonctionnement de la partie logique. C'est pourquoi la mémoire qui s'intègre au BEOL est avantageuse car son intégration est faite en fin de procédé de fabrication, donc moins perturbante pour la partie logique. Cependant ce type de technologie est loin d'être mature, incitant les principaux acteurs de la microélectronique à poursuivre le développement des cellules mémoires de petite taille vers une intégration à stockage de charges. L'adaptation des nœuds technologiques inférieurs à 28nm reste un challenge majeur pour les prochaines années.

III. Les limites des réductions dimensionnelles

Parmi les différents concepts de mémoire à stockage de charges, nous avons sélectionné la Flash NOR. Cette eNVM, est une candidate sérieuse au développement des prochaines générations de mémoires à stockage de charges pour les nœuds technologiques 40nm et 28nm. Nous développerons les enjeux dimensionnels de la cellule, où sont les gains de densité potentiels et quelles sont les limites apparentes.

Pour garantir la compatibilité avec l'intégration de la logique sur des nœuds avancés, cette nouvelle cellule est développée sur un substrat UTBB FD-SOI (Ultra Thin Body Box Fully Depleted Silicon On Insulator), représentée sur la Figure 4. Ce processus est basé sur l'intégration d'une fine couche d'isolant (Buried Oxide) au-dessus du silicium, et d'un film de Silicium qui constitue le canal du transistor sans nécessité de dopage. Avec cette architecture, le transistor se trouve dans un régime de déplétion profonde, dépourvu de courant de fuite dans le substrat, le rendant alors plus performant.

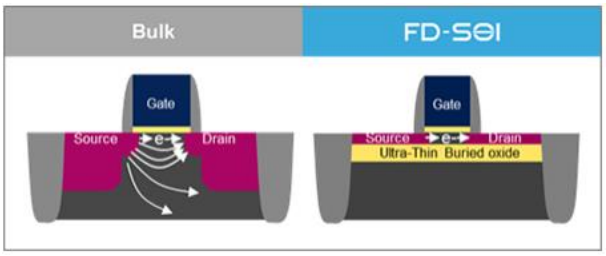


Fig.4. Architecture FD-SOI sur un transistor

Cette architecture permet la réduction de la taille de la cellule sans impacter son fonctionnement et sa performance, néanmoins la réduction dimensionnelle peut être limitée par la technologie. Pour mieux comprendre ces limites, prenons une architecture Flash NOR représentée sur la Figure 5.

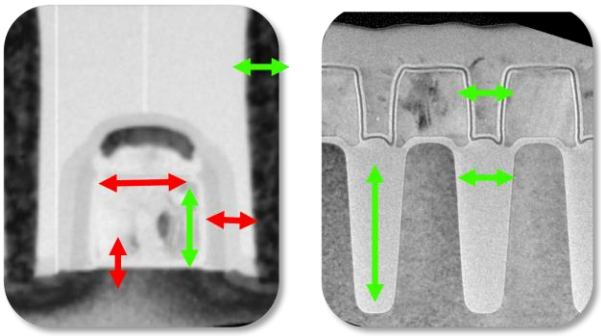


Fig.5. Limites (rouge) et améliorations (vert) dimensionnelles sur une cellule Flash NOR (coupe en L à gauche, coupe en W à droite)

Pour réduire les dimensions de la cellule, certains paramètres ne peuvent être modifiés, sous peine de générer des effets secondaires. C'est le cas de l'épaisseur d'oxyde se trouvant sous la grille flottante. En effet, si nous réduisons la couche d'oxyde tunnel, un courant de SILC (Stress Induced Leakage Current) se crée par la création de charges positives dans l'oxyde qui permettent aux électrons de traverser l'oxyde tunnel par un mécanisme tunnel assisté par des pièges, ou par la génération de charges positives qui abaissent la barrière tunnel à franchir. Les phases d'écriture/effacement génèrent des pièges dans le volume de l'isolant qui sont à l'origine de ce courant de SILC, augmentant la fuite de la charge stockée dans la grille flottante et donc accélère la perte de l'information. De même, si l'on souhaite diminuer la distance entre le contact sur le drain et la grille, une forte tension durant l'effacement est générée. La réduction de la longueur de grille engendre également un effet appelé DIBL (Drain Induced Barrier Lowering), correspondant à la création d'un canal court abaissant la hauteur de barrière de potentiel à l'interface et augmente alors la consommation de la cellule. Ces trois paramètres dimensionnels ne sont donc pas modifiables pour réduire la cellule.

Cependant, il est possible de réduire la taille de la cellule en diminuant la taille du contact, la distance entre les zones actives, la profondeur du STI (Shallow Trench Isolation) ou encore l'épaisseur du Si poly de la grille flottante. Modifier la distance entre les *actives* nécessite d'être vigilant sur le gain de la cellule en lecture et sur ses performances en effacement. En ce qui concerne la profondeur du STI, sa modification entraine la diminution de sa largeur lors du processus de fabrication et de son rôle d'isolation. Pour ce qui est de l'épaisseur du Si poly, tout comme le STI, sa largeur nécessite d'être réduite. Mais il est d'autant plus intéressant d'observer cette répercussion sur la cellule, en

particulier sur le facteur de couplage et le nombre d'électrons stockés dans la grille que nous allons modéliser.

IV. Modélisation

Une cellule Flash NOR possède, comme toute cellule mémoire, deux états avec deux quantités de charges correspondant à deux états logiques '0' et '1'. Dans ce cas, nous parlons des états effacé et programmé de la mémoire, dont la quantité de charges modifie la tension seuil du transistor qui le caractérise. Nous obtenons alors une fenêtre de programmation telle que :

$$\Delta V_t = V_{tprog} - V_{terase} \quad [1]$$

Si nous observons le réseau capacitif représenté sur la Figure 6, nous pouvons calculer la quantité de charges nécessaire dans la grille flottante, permettant de discriminer l'état '0' de l'état '1' :

$$Q_{fg} = C_{pp}(V_{fg} - V_{cg}) + C_d(V_{fg} - V_d) + C_b(V_{fg} - V_b) + C_s(V_{fg} - V_s) \quad [2]$$

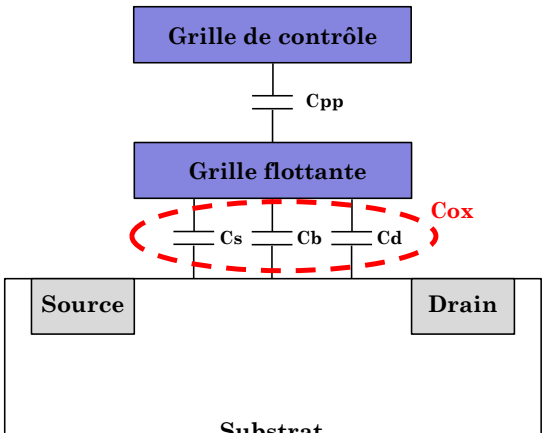


Fig.6. Réseau capacitif d'une cellule Flash NOR

La capacité totale de la cellule s'exprime par :

$$C_{tot} = C_{pp} + C_d + C_s + C_b \quad [3]$$

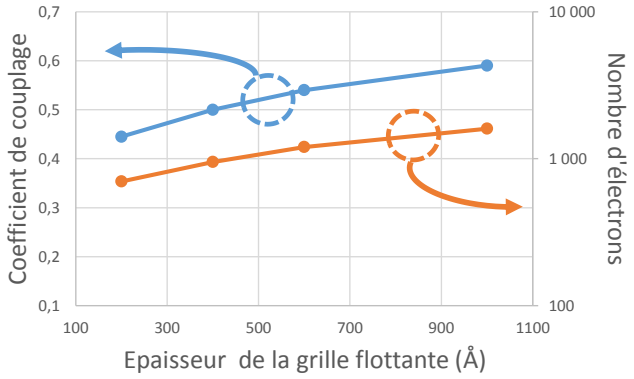
Avec l'équation [3], nous pouvons alors déterminer le coefficient de couplage qui évalue la performance de la cellule, exprimé en pourcentage :

$$\alpha_g = \frac{C_{pp}}{C_{tot}} \quad [4]$$

En connaissant la quantité de charges dans la grille flottante, il est intéressant de calculer le nombre d'électrons qu'elle stocke. Avec [2] et la quantité de charges d'un électron, nous avons :

$$n = \frac{Q_{fg}}{Q_e} \quad [5]$$

Avec [4] et [5], nous pouvons observer sur la Figure 7 l'impact de la modification de l'épaisseur de Si poly de la grille flottante sur le fonctionnement de la cellule. En diminuant cette épaisseur, le nombre d'électrons et le facteur de couplage diminuent également. Cette modélisation, qui tend vers le nœud technologique 28nm, permet de trouver un compromis entre les réductions dimensionnelles et l'efficience de la cellule.



~~Fig.8.~~Fig.7. Modélisation du coefficient de couplage et du nombre d'électrons d'une cellule Flash NOR en fonction de l'épaisseur de la grille flottante

Dans notre cas, il est conseillé de ne pas obtenir moins de 1.000 électrons pour une raison de fiabilité. En effet, si le nombre d'électrons est faible, la perte d'un électron aura plus d'impact sur le fonctionnement de la cellule. En ce qui concerne le coefficient de couplage, sa valeur doit être comprise entre 0,6 et 0,7. Ce facteur indique l'influence du potentiel de la grille de contrôle sur le potentiel de la grille flottante. Si sa valeur est faible, les modes d'écriture et d'effacement nécessiteront des tensions élevées afin d'obtenir la même valeur du potentiel flottant. Si au contraire la valeur du facteur de couplage est élevée, cela signifie que l'influence de la charge de la grille flottante devient négligeable devant la valeur du potentiel du point flottant. Ceci se traduit par la diminution de la fenêtre de programmation (différence entre les tensions seuils des modes d'écriture et d'effacement), rendant alors la distinction entre l'état écrit et effacé à la lecture difficile. Dû à l'enjeu de la miniaturisation, il est nécessaire de conserver une valeur du coefficient de couplage acceptable pour le bon fonctionnement de la cellule.

V. Conclusion

Les mémoires non volatiles ne cessent d'évoluer. Les nombreuses applications de ces mémoires embarquées imposent des critères d'utilisation, devenant alors des freins technologiques. Le développement des nœuds avancés est aujourd'hui un défi industriel majeur. Malgré les nombreuses limitations, des alternatives innovantes sont possibles, présentant des résultats encourageants. La variation de l'épaisseur de la grille flottante ou encore du STI ont montré une modification de la cellule Flash NOR tant sur le plan dimensionnel mais surtout sa performance, laissant envisageable de nombreuses pistes pour l'évolution des mémoires vers des nœuds technologique avancées.

Remerciements

Les auteurs remercient le GIP-CNFM pour le prix du meilleur poster, relatif à ces travaux, attribué lors des Journées Nationales du Réseau Doctoral en Micro-nanoélectronique 2016.

Références

1. J.-U. Han, Y.K. Lee, C.M. Jeon, J. Ryu, *Both NOR and NAND embedded hybrid Flash for S-SIM application using 90nm process technology*, Samsung, IEEE International Memory Workshop (2009).
2. L. Masoero, G. Molas, F. Brun, M. Gély, *Scalability of split-gate charge trap memories down to 20nm for low-power embedded memory*, IEEE International (2011).
3. H.Guan, D.Lee, G.P.Li, *An analytical model for optimization of programming efficiency and uniformity of split gate source-side injection SuperFlash memory*, IEEE Transactions on electron devices, vol. 50, no. 3 (2003).
4. *SuperFlash EEPROM technology*, SST, Technical paper (2001).
5. M. Terai, Y. Tsuji, S. Kotsuji, S. Fujieda, *Trapped-hole-enhanced erase-level shift by FN-stress disturb in sub-90-nm-node embedded SONOS memory*, IEEE transactions on electron devices, vol. 55, no. 6 (2008).
6. T. Ren, L. Pan, Z. Liu, J. Zhu, *A novel single poly EEPROM cell structure on thin oxide tunnel technology,* Solid state and integrated circuits technology, vol. 1 (2004).
7. B. Eitan, D. Frohman-Bentchkowsky, *Hot-electron injection into the oxide in n-channel MOS devices*, IEEE transactions on electron devices, vol. 28, pp. 328-340 (1981).
8. R.H. Fowler, L. Nordheim, *Electron emission in intense electric fields*, Proceedings of the royal society of London, Series A, vol. 119, no. 781, pp. 173-181 (1928).
9. Y.K. Lee, J.H. Moon, Y.H. Kim, M.-J. Chun, *2T-FN eNVM with 90 nm logic process for smart card*, Non-volatile semiconductor memory workshop (2008).
10. E.I. Vatajelu, H. Aziza, C. Zambelli, *Nonvolatile memories: present and future challenges*, 9th international design and test symposium (2014).