Optimisation et réduction de la variabilité d'une nouvelle architecture mémoire non volatile ultra basse consommation

El Amine Agharben

N°d'ordre NNT : ***2017LYSEM013***

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de
l'Ecole des Mines de Saint-Etienne

Ecole Doctorale N° 488
Sciences, Ingénierie, Santé

Discipline : Microélectronique

Soutenue publiquement le 05/05/2017, par :

El Amine AGHARBEN

Optimisation et réduction de la variabilité d'une nouvelle architecture mémoire non volatile ultra basse consommation

---

Devant le jury composé de :

GHIBAUDO Gerard / Directeur de recherche CNRS / IMEP-LAHC	Rapporteur
OUSSAR Yacine / Maître de conférences - HDR / ESCPI Paris	Rapporteur
MULLER Christophe / Professeur - Directeur/ CNRS Grenoble	Examinateur
REIS MARCO / Professeur / Université Coimbra (Portugal)	Examinateur
SERGENT Michelle / Professeur / Aix-Marseille université	Examinatrice
ROUSSY Agnès / Maître de conférences – HDR / EMSE CMP-SGC	Directrice de thèse
BOCQUET Marc / Maître de conférences / IM2NP Aix-Marseille université	Encadrant
BILECI Marco / Ingénieur Device / STMicroelectronics	Encadrant industriel

Spécialités doctorales	Responsables:	Spécialités doctorales	Responsables:
SCIENCES ET GENIE DES MATERIAUX	K. Wolski Directeur de recherche	MATHEMATIQUES APPLIQUEES	O. Roustant, Maître-assistant
MECANIQUE ET INGENIERIE	S. Drapier, professeur	INFORMATIQUE	O. Boissier, Professeur
GENIE DES PROCEDES	F. Gruy, Maître de recherche	IMAGE, VISION, SIGNAL	JC. Pinoli, Professeur
SCIENCES DE LA TERRE	F. Gruy, Maître de recherche	GENIE INDUSTRIEL	X. Delorme, Maître assistant
SCIENCES ET GENIE DE L'ENVIRONNEMENT	D. Graillot, Directeur de recherche	MICROELECTRONIQUE	Ph. Lalevée, Professeur

EMSE : Enseignants-chercheurs et chercheurs autorisés à diriger des thèses de doctorat (titulaires d'un doctorat d'État ou

ABSI	Nabil	CR	Génie industriel	CMP
AUGUSTO	Vincent	CR	Image, Vision, Signal	CIS
AVRIL	Stéphane	PR2	Mécanique et ingénierie	CIS
BADEL	Pierre	MA(MDC)	Mécanique et ingénierie	CIS
BALBO	Flavien	PR2	Informatique	FAYOL
BASSEREAU	Jean-François	PR	Sciences et génie des matériaux	SMS
BATTON-HUBERT	Mireille	PR2	Sciences et génie de l'environnement	FAYOL
BEIGBEDER	Michel	MA(MDC)	Informatique	FAYOL
BLAYAC	Sylvain	MA(MDC)	Microélectronique	CMP
BOISSIER	Olivier	PR1	Informatique	FAYOL
BONNEFOY	Olivier	MA(MDC)	Génie des Procédés	SPIN
BORBELY	Andras	MR(DR2)	Sciences et génie des matériaux	SMS
BOUCHER	Xavier	PR2	Génie Industriel	FAYOL
BRODHAG	Christian	DR	Sciences et génie de l'environnement	FAYOL
BRUCHON	Julien	MA(MDC)	Mécanique et ingénierie	SMS
BURLAT	Patrick	PR1	Génie Industriel	FAYOL
CHRISTIEN	Frédéric	PR	Science et génie des matériaux	SMS
DAUZERE-PERES	Stéphane	PR1	Génie Industriel	CMP
DEBAYLE	Johan	CR	Image Vision Signal	CIS
DELAFOSSE	David	PR0	Sciences et génie des matériaux	SMS
DELORME	Xavier	MA(MDC)	Génie industriel	FAYOL
DESRAYAUD	Christophe	PR1	Mécanique et ingénierie	SMS
DJENIZIAN	Thierry	PR	Science et génie des matériaux	CMP
DOUCE	Sandrine	PR2	Sciences de gestion	FAYOL
DRAPIER	Sylvain	PR1	Mécanique et ingénierie	SMS
FAVERGEON	Loïc	CR	Génie des Procédés	SPIN
FEILLET	Dominique	PR1	Génie Industriel	CMP
FOREST	Valérie	MA(MDC)	Génie des Procédés	CIS
FOURNIER	Jacques	Ingénieur chercheur CEA	Microélectronique	CMP
FRACZKIEWICZ	Anna	DR	Sciences et génie des matériaux	SMS
GARCIA	Daniel	MR(DR2)	Sciences de la Terre	SPIN
GAVET	Yann	MA(MDC)	Image Vision Signal	CIS
GERINGER	Jean	MA(MDC)	Sciences et génie des matériaux	CIS
GOEURIOT	Dominique	DR	Sciences et génie des matériaux	SMS
GONDRAN	Natacha	MA(MDC)	Sciences et génie de l'environnement	FAYOL
GRAILLOT	Didier	DR	Sciences et génie de l'environnement	SPIN
GROSSEAU	Philippe	DR	Génie des Procédés	SPIN
GRUY	Frédéric	PR1	Génie des Procédés	SPIN
GUY	Bernard	DR	Sciences de la Terre	SPIN
HAN	Woo-Suck	MR	Mécanique et ingénierie	SMS
HERRI	Jean Michel	PR1	Génie des Procédés	SPIN
KERMOUCHE	Guillaume	PR2	Mécanique et Ingénierie	SMS
KLOCKER	Helmut	DR	Sciences et génie des matériaux	SMS
LAFOREST	Valérie	MR(DR2)	Sciences et génie de l'environnement	FAYOL
LERICHE	Rodolphe	CR	Mécanique et ingénierie	FAYOL
MALLIARAS	Georges	PR1	Microélectronique	CMP
MOLIMARD	Jérôme	PR2	Mécanique et ingénierie	CIS
MOUTTE	Jacques	CR	Génie des Procédés	SPIN
NIKOLOVSKI	Jean-Pierre	Ingénieur de recherche	Mécanique et ingénierie	CMP
NORTIER	Patrice	PR1		SPIN
OWENS	Rosin	MA(MDC)	Microélectronique	CMP
PERES	Véronique	MR	Génie des Procédés	SPIN
PICARD	Gauthier	MA(MDC)	Informatique	FAYOL
PIJOLAT	Christophe	PR0	Génie des Procédés	SPIN
PIJOLAT	Michèle	PR1	Génie des Procédés	SPIN
PINOLI	Jean Charles	PR0	Image Vision Signal	CIS
POURCHEZ	Jérémy	MR	Génie des Procédés	CIS
ROBISSON	Bruno	Ingénieur de recherche	Microélectronique	CMP
ROUSSY	Agnès	MA(MDC)	Microélectronique	CMP
ROUSTANT	Olivier	MA(MDC)	Mathématiques appliquées	FAYOL
STOLARZ	Jacques	CR	Sciences et génie des matériaux	SMS
TRIA	Assia	Ingénieur de recherche	Microélectronique	CMP
VALDIVIESO	François	PR2	Sciences et génie des matériaux	SMS
VIRICELLE	Jean Paul	DR	Génie des Procédés	SPIN
WOLSKI	Krzystof	DR	Sciences et génie des matériaux	SMS
XIE	Xiaolan	PR1	Génie industriel	CIS
YUGMA	Gallian	CR	Génie industriel	CMP

TABLE DES MATIERES

TABLE DES FIGURES

GLOSSAIRE

AHM	:	Ashable Hard Mask
BARC	:	Bottom Anti Reflective Coating
BEOL	:	Back End Of Line
BL	:	Bit Line
CG	:	Control Gate
CHE	:	Channel Hot Electron
CMOS	:	Complementary Metal Oxide Semiconductor
CMP	:	Chemical Mechanical Polishing
CVD	:	Chemical Vapor Deposition
DIBL	:	Drain Induced Barrier Leackage
DOE	:	Design Of Experiments
DRAM	:	Dynamic Random Access Memory
EEPROM	:	Electrically Erasable Programmable Read Only Memory
EG	:	Erase Gate
EPROM	:	Erasable Programmable Read Only Memory
eSTM	:	Embedded Select Trench Memory
EWS	:	Electrical Wafer Sort
FEOL	:	Front End Of Line
FG	:	Floating Gate
FN	:	Fowler-Nordheim
FRAM	:	Ferroelectric Random Access Memory
GDS	:	Graphic Database System
HDP	:	High Density Plasma
HHI	:	Hot Holes Injection
HMDS	:	Hexamethyldisilazane
HRS	:	High Resistance State
IPD	:	Image Plan Deviation
ISSG	:	In Situ Steam Generated
ITRS	:	International Technology Roadmap for Semiconductors
LER	:	Line Edge Roughness
LRS	:	Low Resistance State
LWR	:	Line Width Roughness
MEB	:	Microscope Electronique à Balayage
MIM	:	Métal Isolant Métal
MIMO	:	Multiple Input Multiple Output
MISO	:	Multiple Input Single Output
MRAM	:	Magnetic Random Access Memory
NFARL	:	Nitride Free Anti Reflective Layer
ONO	:	Oxide Nitride Oxide

OPC	:	Optical Proximity Correction
OTV	:	Oxide Thickness Variations
PCM	:	Phase Change Memory
PLS	:	Partial Least Squares
PROM	:	Programmable Read Only Memory
R2R	:	Run-to-Run
RDF	:	Random Dopant Fluctuations
ReRAM	:	Resistive Random Access Memory
ROM	:	Random Access Memory
RSC	:	Reticle Shape Correction
RSU	:	Remote-sense & Switch Unit
SCE	:	Short Channel Effect
SEM	:	Scanning Electron Microscopy
SEMCD	:	Scanning Electron Microscopy Critical Dimension
SG	:	Select Gate
SILC	:	Stress Induced Leackage Current
SIMO	:	Single Inputs Multiple Outputs
SIMS	:	Spectrométrie de Masse à Ionisation Secondaire
SISO	:	Single Inputs Single Outputs
SMU	:	Source Measure Unit
SPGU	:	Semiconductor Pulse Generator Unit
SRAM	:	Static Random Access Memory
SSI	:	Source Side Injection
STI	:	Shallow Trench Isolation
TCAD	:	Technology Computer Aided Design
TEM	:	Transmission Electron Microscopy
WGFMU	:	Waveform Generator / Fast Measurement Unit
WL	:	Word Line

INTRODUCTION GENERALE

Le marché mondial des semi-conducteurs connait une croissance continue due à l'essor de l'électronique grand public et entraîne dans son sillage le marché des mémoires non volatiles. L'importance de ces produits mémoires est accentuée depuis le début des années 2000 par la mise sur le marché de produits nomades tels que les smartphones ou plus récemment les produits de l'internet des objets. De par leurs performances et leur fiabilité, la technologie Flash constitue, à l'heure actuelle, la référence en matière de mémoire non volatile. Cependant, ces mémoires étant en passe d'atteindre leurs limites de miniaturisation, plusieurs dispositifs alternatifs sont actuellement envisagés par les industriels du secteur, de manière à anticiper les demandes du marché ces prochaines années. De plus la maîtrise des différentes étapes du processus de fabrication des circuits CMOS (Complementary Metal Oxide Semiconductor) est une priorité pour les différents acteurs de ce secteur. Dans un environnement de production, les procédés de fabrication souffrent d'une fluctuation permanente. Cette variation est engendrée par plusieurs facteurs : la fluctuation des propriétés de la matière première (résine photosensible, substrat…) et l'évolution des équipements dans le temps (vieillissement, changement des flux de gaz…).

Parallèlement, le coût élevé des équipements en microélectronique rend impossible leur amortissement sur une génération technologique, d'autant plus que la durée de vie des générations décroît rapidement. Ceci incite l'industriel à adapter des équipements d'ancienne génération à des procédés de fabrication plus exigeants. Cette stratégie n'est pas sans conséquence sur la dispersion des caractéristiques physiques (dimension géométrique, épaisseur…) et électriques (courant, tension…) des dispositifs. Dans ce contexte, le sujet de ma thèse est d'optimiser et de réduire la variabilité d'une nouvelle architecture mémoire non volatile ultra basse consommation. Ce travail de thèse a été réalisé dans le cadre d'une convention CIFRE en collaboration entre la société STMicroelectronics sur le site de Rousset, le Centre Microélectronique de Provence et l'Institut Matériaux Microélectronique Nanosciences de Provence (IM2NP). Ce manuscrit de thèse sera présenté en 4 chapitres.

Le premier chapitre commence par présenter le contexte économique et les perspectives du marché des mémoires non volatiles. La partie suivante décrit leurs principes de fonctionnement et les limites physiques de leur miniaturisation. Et pour finir une description de la nouvelle architecture mémoire de STMicroelectronics dédiée aux applications basse consommation sera faite.

Dans le second chapitre, nous étudierons la nouvelle architecture mémoire appelée eSTM (embedded Select Trench Memory). Cette mémoire aura pour but d'adresser les applications basse consommation afin de concurrencer les mémoires Split Gate et 2T. Dans la première partie nous présenterons le procédé de fabrication de cette architecture et son principe de fonctionnement. Par la suite, les principales caractéristiques électriques de la cellule mémoire eSTM seront présentées. Enfin une comparaison entre les performances de cette nouvelle cellule avec une mémoire Flash standard sera présentée afin de mettre en avant le gain en consommation.

Le troisième chapitre est dédié à l'étude de variabilité des étapes du procédé de fabrication de la cellule eSTM. Le premier objectif est de trouver une méthodologie statistique capable d'identifier les caractéristiques physiques qui sont à l'origine de la variabilité des paramètres électriques. La seconde partie sera consacrée à une étude approfondie des criticités du procédé de fabrication et leurs impacts sur les performances électriques.

Le dernier chapitre traite des différentes optimisations du procédé de fabrication de la cellule eSTM afin de corriger les différentes variabilités trouvées dans le Chapitre 3. La première partie présentera la description du nouveau procédé de gravure de l'empilement mémoire et son impact sur les performances électriques de la cellule eSTM. Dans la partie suivante, les étapes de la mise en place d'une boucle de régulation entre la gravure de la tranchée du transistor de sélection et celle de l'empilement mémoire seront décrites. Et pour finir nous présenterons l'impact de l'optimisation du procédé de fabrication de la cellule eSTM sur sa fiabilité.

Chapitre 1 : ETAT DE L'ART DES MEMOIRES NON VOLATILES

Sommaire

I. Introduction

Depuis l'invention du premier circuit intégré en 1958, la microélectronique a connu un essor économique considérable. Le transistor qui a vu le jour en 1947, est la brique de base des circuits intégrés. Les performances de ces circuits n'ont cessé de croître grâce à la miniaturisation des transistors dictée par la loi de Moore [1]. Gordon Moore avait énoncé en 1965 une loi très simple qui prédisait que sur une surface fixée, le nombre de composants pouvant être intégrés devrait doubler tous les 18 mois. Or dans un futur proche, pour les dispositifs à semi-conducteur, cette loi sera confrontée à une barrière technologique infranchissable.

Ce chapitre présente, dans un premier temps, le contexte économique, l'environnement technique et le principe de fonctionnement des mémoires à stockages de charges. Ensuite, nous verrons les limitations physiques liées à la réduction de la taille des cellules mémoires ainsi que les principales solutions envisagées. Et en particulier la nouvelle architecture mémoire développée à STMicroelectronics sur le site de Rousset. L'étude du procédé de fabrication de cette architecture ainsi que les différents tests électriques seront développées tout au long de ce manuscrit.

II. L'industrie du semi-conducteur

1. Le marché des mémoires non volatiles

Le marché mondial des semi-conducteurs connaît une croissance soutenue depuis la fin des années 80 (Figure 1.1). En dépit des crises financières qui l'ont frappée au début des années 2000 (éclatement de la bulle spéculative « internet ») ou des incertitudes économiques qui planent actuellement sur le marché des smartphones et des tablettes, cette industrie affiche à l'heure actuelle un revenu mondial annuel proche de 340 milliards de dollars, soit 6 fois plus qu'au début des années 90. L'extrême vivacité de ce secteur industriel tient à la fois à sa capacité d'infiltrer un nombre considérable de marchés, allant de l'industrie de l'automobile à l'électronique grand public, mais aussi aux efforts constants des concepteurs de systèmes électroniques en matière d'innovation technologique.

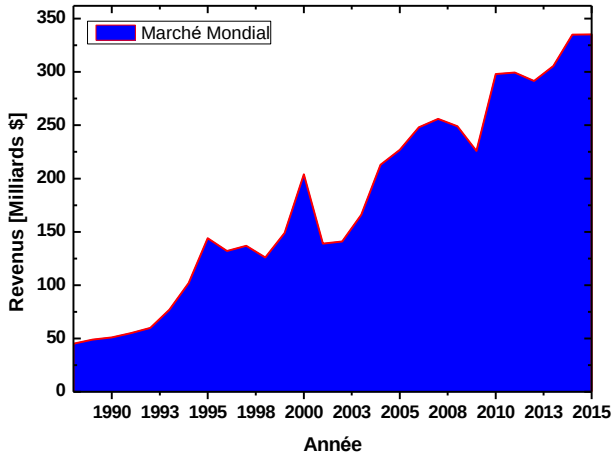


Figure 1.1 – Revenus mondiaux de l'industrie du semi-conducteur depuis 1988

(Source : www.statistica.com)

À titre d'illustration, dans le secteur de l'électronique grand public, l'introduction récente de l'internet des objets, associée à l'essor spectaculaire des réseaux sociaux et du big data, semblent aujourd'hui indiquer que les grands développeurs de systèmes électroniques (Apple, Google, Samsung …) sont en mesure de renouveler la demande des consommateurs pour la décennie à venir. Outre le nombre de transistors par puce, dont la densité au centimètre carré est pilotée depuis la fin des années 70 par la « Loi de Moore », la quantité de mémoire embarquée constitue un autre indicateur de la performance d'un système électronique. La Figure 1.2 montre l'évolution des revenus financiers générés depuis 2013 par les différents types de mémoires de l'industrie du semi-conducteur. Elle met en évidence que le marché des mémoires va bénéficier d'une croissance dans les années à venir grâce à l'engouement que suscite l'internet des objets.

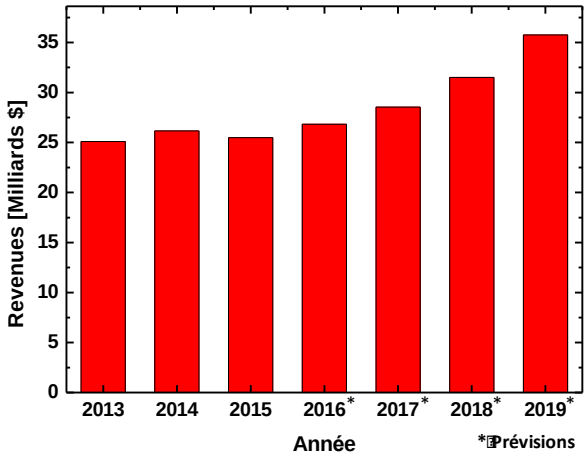


Figure 1.2 : Revenus mondiaux des mémoires à semi-conducteur depuis 2013

(Source : www.statistica.com)

2. Classification des mémoires

Afin de conserver des données informatiques, il existe différents types de mémoire avec des performances adaptées selon l'application. En effet, certaines applications nécessitent des temps d'écriture ou de lecture faibles tandis que d'autres devront être capables de stocker beaucoup de données. Les caractéristiques permettant de classifier les différents types de mémoire sont :

- Endurance
- Consommation
- Rapidité en écriture/lecture
- Coût de production
- Rétention
- Taille ou densité

Aucune mémoire actuelle n'étant capable de conjuguer tous les avantages, la mémoire parfaite n'existe pas. Ainsi il sera nécessaire d'effectuer des compromis sur chacune de ces caractéristiques élémentaires pour une adéquation optimale de la mémoire avec son application.

De façon générale, on distingue les mémoires en deux catégories, les mémoires volatiles et non volatiles (Figure 1.3).

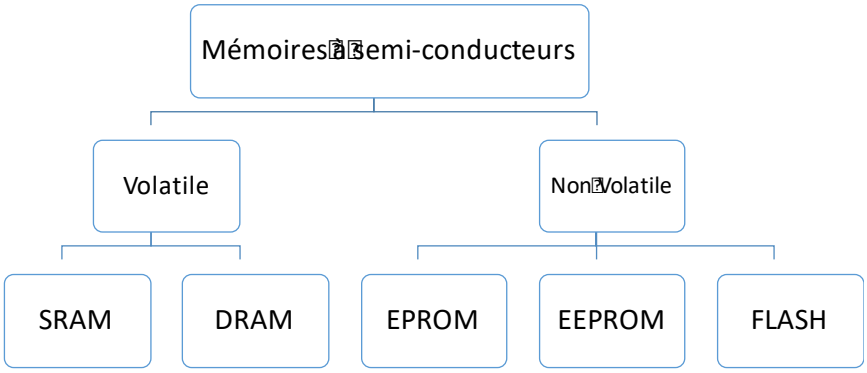


Figure 1.3 - Arborescence des mémoires à semi-conducteurs

Les mémoires volatiles sont des mémoires rapides utilisées pour le stockage de l'information. Leurs vitesses de programmation élevées font d'elles des mémoires largement utilisées dans les mémoires caches et centrales, et ce malgré le fait que l'information stockée soit perdue en cas d'une coupure d'alimentation.

Ce type de mémoire se partage en deux familles :

- Les mémoires statiques, ou SRAM (Static Random Access Memory) [2], [3], sont basées sur des bascules électroniques, typiquement à six transistors, comme illustré en Figure 1.4. La terminologie statique indique qu'elles n'ont pas besoin d'être reprogrammées régulièrement, de plus elles présentent des vitesses de programmation et d'accès extrêmement rapides. En revanche le coût, la taille et la consommation restent leurs défauts majeurs.

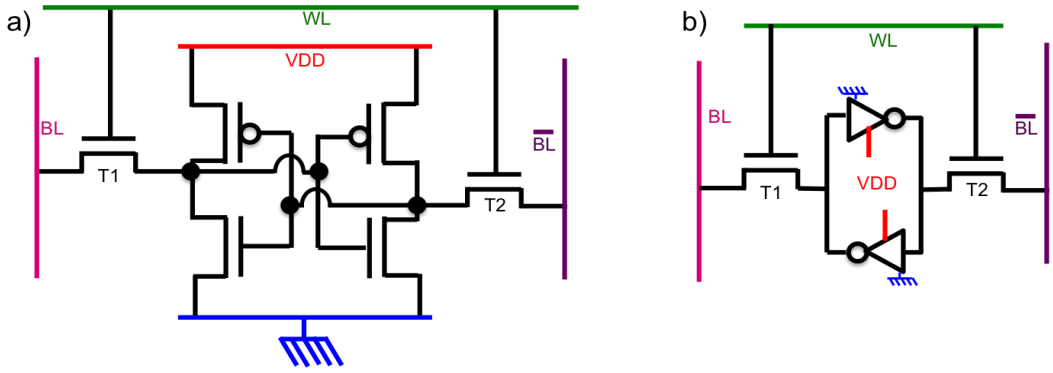


Figure 1.4 - a) Schéma électrique b) Schéma logique d'une SRAM

- Les mémoires dynamiques, ou DRAM (Dynamic Random Access Memory) [4], sont basées sur la combinaison d'une capacité et d'un transistor, comme présenté en Figure 1.5. Elles sont capables de stocker l'information pendant quelques millisecondes seulement. Afin de garder l'information de ce type de mémoire, un rafraîchissement doit

être fait régulièrement. Il consiste en une lecture de la part du contrôleur mémoire suivi d'une réécriture. Ces mémoires sont principalement utilisées comme mémoire centrale, car elles présentent une densité très supérieure, et donc un coût attractif comparé aux technologies SRAM.

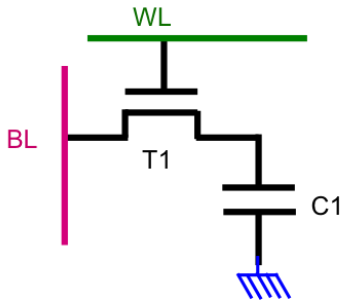


Figure 1.5 - Schéma électrique d'une DRAM

Les mémoires non volatiles sont capables de garder l'information stockée en l'absence d'alimentation électrique. Dans cette thèse, l'étude va se concentrer sur les mémoires non volatiles à stockage de charges qui constituent un sous-groupe des mémoires à semi-conducteurs. Cependant il est important de définir les différents dispositifs qui ont permis le stockage de l'information :

- La ROM, (Read Only Memory), est une mémoire programmée au cours de sa fabrication [5], [6]. Une fois programmée, il sera impossible d'effacer ou de changer son état. Elles sont souvent utilisées pour stocker les informations nécessaires au démarrage d'un ordinateur (BIOS, microcode…).
- La PROM, (Programmable Read Only Memory), est une mémoire similaire à la ROM. Cependant, la phase de programmation est faite par l'utilisateur. Elle a été inventée en 1956 et constitue une alternative moins onéreuse que la mémoire ROM, car elle n'a pas besoin d'un nouveau masque pour la programmation.
- L'EPROM, (Erasable Programmable Read Only Memory) [7], est une mémoire qui peut être programmée et effacée par l'utilisateur. La particularité de ce dispositif est la présence d'une grille flottante emprisonnée entre deux oxydes : un côté canal permettant la programmation, nommé "oxyde tunnel", et du côté grille de contrôle assurant le couplage électrostatique, nommé "oxyde d'interpoly". L'écriture se fait en injectant des électrons dans la grille flottante à travers l'oxyde tunnel. L'effacement s'effectue en l'exposant aux rayons UV permettant de fournir assez d'énergie aux électrons piégés pour quitter la grille flottante.
- L'EEPROM, (Electrically Erasable Programmable Read Only Memory) [8], [9], est basée sur le principe l'EPROM mais offre la possibilité d'un effacement électrique. Ce dispositif permet une modification aisée du contenu mémoire ainsi qu'un accès et une programmation bit à bit grâce au transistor de sélection en série de chaque transistor mémoire.
- La mémoire Flash, [10], [11] présente un empilement similaire à la mémoire EEPROM, mais le transistor de sélection est supprimé pour augmenter la densité du plan mémoire. Historiquement, le nom vient de son effacement par page. En fonction de leurs applications, les mémoires flash peuvent être utilisées dans deux

architectures différentes. D'une part la mémoire type NAND, dont le matriçage extrêmement dense et la faible vitesse en accès aléatoire la destine au stockage massif de données ; et d'autre part, la mémoire de type NOR dont la rapidité et la fiabilité permettent le stockage et l'exécution en temps réel du microcode des systèmes embarqués. Ces architectures seront présentées en détail dans la suite de ce chapitre.

3. L'architecture des mémoires Flash

Les mémoires Flash sont organisées en réseaux de lignes (Word Lines ou WL) et de colonnes (Bit Lines ou BL). L'architecture du réseau est déterminée par le type de connexion (Figure 1.6).

- NOR : l'architecture NOR a été introduite pour la première fois par Intel en 1988. Les cellules sont connectées en parallèle. Les grilles sont reliées entre elles par l'intermédiaire de la Word Line, tandis que le drain est partagé le long de la Bit Line. Généralement la programmation se fait par injection d'électron chaud (Channel Hot Electron CHE) et l'effacement par Fowler-Nordheim (FN). Le fait que le drain de chaque cellule peut être sélectionné de manière individuelle permet un accès aléatoire d'une cellule dans le plan mémoire. Néanmoins la présence des contacts de drain pour chaque cellule mémoire limite la miniaturisation[1] à $6F^2$. Une lecture rapide, une bonne fiabilité et un mécanisme d'écriture relativement rapide font de l'architecture NOR la technologie la plus appropriée pour les applications embarquées. La cellule mémoire étudiée dans ce manuscrit sera intégrée dans une architecture NOR pour des applications embarquées ultra basse consommation.
- NAND : l'architecture NAND a été présentée par Toshiba en 1987. Les mémoires Flash NAND ont été baptisées ainsi, car les cellules sont connectées en série, tout comme les portes logiques NAND. Dans cette architecture, les cellules sont connectées en série. Les grilles de contrôles sont reliées les unes aux autres, mais les drains ne sont pas adressables individuellement. En l'absence des contacts de drain le seul mécanisme de programmation possible est de type Fowler-Nordheim. L'un des atouts majeurs de cette technologie est son faible coût de fabrication par bit, notamment grâce à l'absence de contact de drain. En revanche, cette absence ne permet pas l'accès aléatoire rapide, ce qui est rédhibitoire pour l'exécution d'un microcode embarqué. Il faut en effet, charger l'intégralité d'une ligne de bits pour lire l'état d'une cellule individuelle. Pour cette raison, le domaine d'application de la Flash NAND est le stockage massif de données : clés USB, cartes mémoires, smartphones, tablettes et PC.

[1] F est la taille minimale

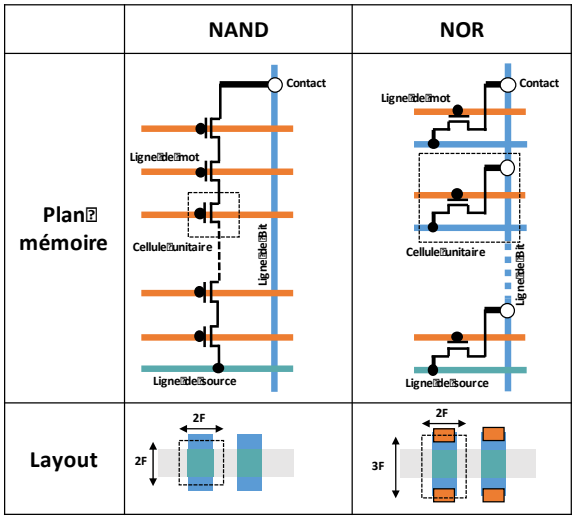


Figure 1.6 - Architecture NAND (gauche) et NOR (droite) du plan mémoire (Source : www.micron.com)

III. Principe de fonctionnement des mémoires à stockage de charges

La compréhension des concepts et fonctionnalités de base des mémoires à stockage de charge est essentielle pour appréhender les travaux de cette thèse. Dans cette partie, nous allons décrire le principe de fonctionnement du transistor mémoire présent dans les architectures EEPROM et Flash (Figure 1.7).

Lorsque la cellule est effacée, il n'y a pas de charge dans la grille flottante, la tension de seuil V_T est faible et vaut une valeur notée V_{TE}. Au contraire, lorsque la mémoire est programmée (ou écrite) la charge injectée est stockée dans la grille flottante et la valeur de la tension de seuil passe à une valeur supérieure notée V_{TP}. Pour connaître l'état de la cellule mémoire (ex. la quantité de charge piégée), il est nécessaire de polariser la grille de contrôle avec une tension de lecture comprise entre V_{TE} et V_{TP}, puis de déterminer si le courant circule à travers le canal (état ON) ou non (état OFF).

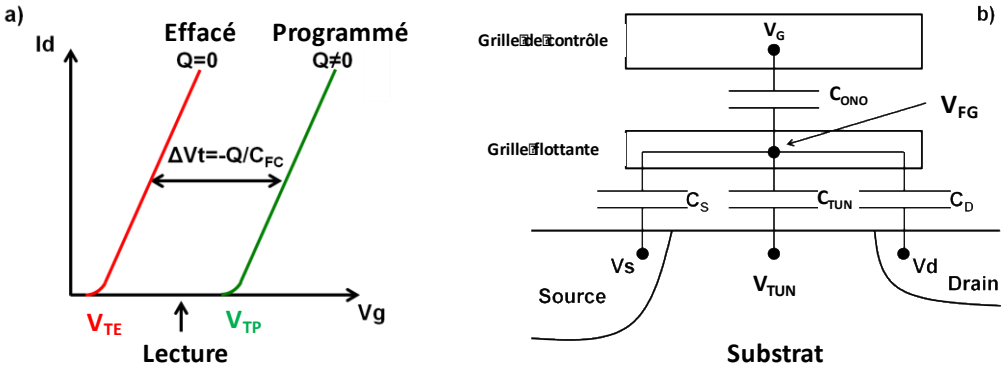


Figure 1.7 - a) Caractéristiques I-V d'un dispositif à grille flottante pour deux valeurs différentes de la charge stockée dans la grille flottante (Q=0 et Q≠0). b) Modèle capacitif d'un transistor à grille flottante

Le transistor à stockage de charges repose sur un transistor MOS avec un niveau de Polysilicium supplémentaire (Figure 1.7b). Les deux niveaux de Polysilicium sont séparés par un diélectrique Oxyde / Nitrure / Oxyde (ONO) nommé « diélectrique interpoly ». Le rôle de cette tri-couche est d'empêcher le passage des charges piégées dans la grille flottante vers la grille de contrôle. Le modèle capacitif (Figure 1.7b) permettra la compréhension du fonctionnement électrique du transistor à stockage de charges. C_{ONO}, C_S, C_D, et C_{TUN} représentent respectivement les capacités entre la grille flottante et la grille de contrôle, la source, le drain et le substrat. Le potentiel de la grille flottante est V_{FG}, V_G est le potentiel de la grille de contrôle, et V_S, V_D, V_B sont respectivement les potentiels de la source, drain et substrat.

1. Structure basique : modèle capacitif

Le fonctionnement de la cellule peut être décrit à l'aide du schéma électrique équivalent présenté dans la Figure 1.7b. Selon ce modèle capacitif, le potentiel de la grille flottante est déterminé par les potentiels des électrodes environnantes et par la charge contenue dans la grille flottante par la relation [12] :

$$V_{FG} = \frac{Q_{FG}}{C_{TOT}} + \alpha_G V_G + \alpha_S V_S + \alpha_D V_D + \alpha_B V_B \quad (1.1)$$

$$avec : \ C_{TOT} = C_{ONO} + C_D + C_S + C_{TUN} \quad (1.2)$$

$$et : \ \alpha_G = \frac{C_{ONO}}{C_{TOT}}, \alpha_S = \frac{C_S}{C_{TOT}}, \alpha_D = \frac{C_D}{C_{TOT}}, \alpha_B = \frac{C_{TUN}}{C_{TOT}} \quad (1.3)$$

Comme l'indique le schéma électrique équivalent de la Figure 1.7b, le canal du transistor est couplé par effet électrostatique à la grille flottante par la capacité C_{TUN}. Par conséquent, l'apparition de la courbe d'inversion est entièrement conditionnée par le potentiel de la grille flottante. Au seuil d'inversion du dispositif, lorsque V_G=V_T, ce potentiel prend la valeur suivante :

$$V_{FG-T} = \frac{Q_{FG}}{C_{TOT}} + \alpha_G V_T + \alpha_S V_S + \alpha_D V_D + \alpha_B V_B \quad (1.4)$$

L'équation 1.4 permet d'expliciter l'évolution de ΔV_T, communément appelé **fenêtre de programmation**, correspondant à l'écart entre la tension de seuil actuelle du dispositif et la tension de seuil pour $\Delta Q_{FG} = 0$, noté V_{T0}.

$$\Delta V_T = V_T - V_{T0} = -\frac{Q_{FG}}{\alpha_G C_{TOT}} = -\frac{Q_{FG}}{C_{ONO}} \quad (1.5)$$

Notons ici que α_G correspond au coefficient de couplage entre la grille de contrôle et la grille flottante. Compte tenu de la définition donnée dans l'équation 1.3, le coefficient de couplage α_G est proportionnel au produit de la permittivité relative du diélectrique interpoly (ε_{r-IPD}) par la surface de recouvrement entre la grille flottante et la grille de contrôle (S_{FG-CG}) :

$$\alpha_G = \frac{C_{ONO}}{C_{TOT}} \propto \varepsilon_{r-IPD} * S_{FG-CG} \quad (1.6)$$

D'après l'ITRS[2] la condition optimale de fonctionnement des transistors à stockage de mémoire est $0{,}6 < \alpha_G < 0{,}7$.

2. Les mécanismes de programmation des mémoires à grille flottante

Dans cette partie nous allons décrire trois mécanismes pour programmer un empilement mémoire à grille flottante : effet Fowler-Nordheim (FN) [13], injection de porteurs chauds (Channel Hot Electron ou CHE)[14] ainsi que l'injection côté source.

- **La programmation par effet Fowler-Nordheim :**

La programmation par effet Fowler-Nordheim consiste en la création d'un courant tunnel entre le canal et la grille flottante à travers l'oxyde de tunnel [15]. Une tension élevée est appliquée sur la grille de contrôle (de l'ordre de 18V) tandis que le substrat est mis à la masse ainsi que la source et le drain (Figure 1.8a). Par effet capacitif la tension de la grille flottante se rapproche des 12V, créant ainsi un champ électrique important aux bornes de l'oxyde tunnel. Ce champ électrique a pour effet de déformer la barrière d'énergie de l'oxyde tunnel autorisant ainsi le passage des électrons de la bande de conduction du canal vers la bande de conduction de la grille flottante (Figure 1.8b). Les électrons sont stockés au fur et à mesure dans la grille flottante réduisant son potentiel jusqu'à ce que le champ électrique aux bornes de l'oxyde tunnel soit insuffisant. Cette méthode de programmation est considérée comme lente (de l'ordre de la milliseconde), mais présente l'avantage d'avoir un courant de programmation négligeable.

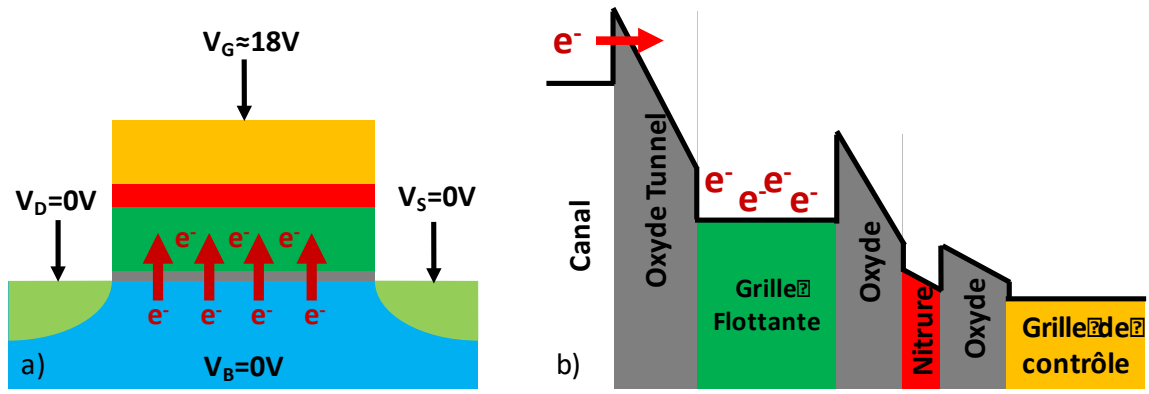


Figure 1.8 - a) Représentation du mécanisme de programmation par Fowler-Nordheim b) Diagramme de bande durant la programmation

- **Programmation par injection de porteurs chauds**

Pour programmer par injection de porteurs chauds, il faut appliquer une tension élevée sur la grille (environ 10V) ainsi que sur le drain (environ 4V) et laisser la source et le substrat à la masse (Figure 1.9). La différence de potentiels entre le drain et la source crée un champ électrique horizontal qui va accélérer les électrons dans la zone de pincement du canal. Certains électrons atteindront une énergie cinétique supérieure à la barrière de l'oxyde tunnel et seront injectés dans la grille flottante grâce au champ électrique vertical dû à la polarisation de la grille de contrôle [16]–[18]. Ce mécanisme de programmation est plus rapide (de l'ordre de la microseconde) que l'injection Fowler-Nordheim. En revanche, seuls quelques électrons parviennent à atteindre la

[2] ITRS : International Technology Roadmap for Semiconductors

grille flottante, laissant les autres traverser le canal. Le courant nécessaire à ce type de programmation est donc important [19].

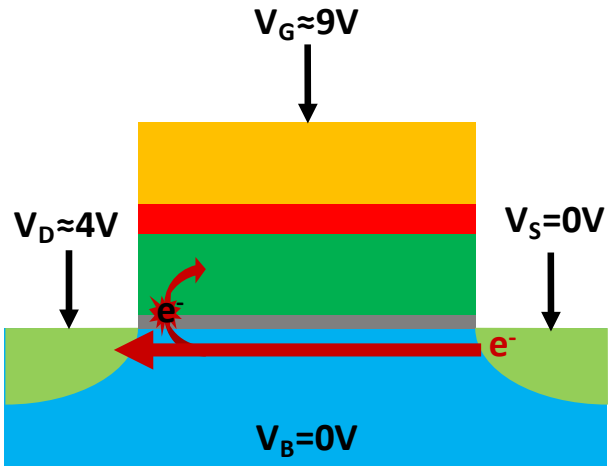


Figure 1.9 - Représentation du mécanisme de programmation par CHE.

- **Programmation par injection côté source (*Source Side Injection*)**

Dans le cas des architectures mémoires à deux transistors (de type split gate), la programmation par injection côté source (*Source Side Injection* ou SSI) est identique au mécanisme CHE, mais réalisée côté source au lieu du côté drain du transistor mémoire. Pour obtenir cette forme de programmation, une architecture composée de deux grilles est nécessaire. L'une aura pour but de contrôler le point mémoire (CG) et l'autre mettra en conduction un transistor de sélection (SG). Le transistor mémoire est polarisé comme dans le cas de la programmation par porteur chaud, c'est à dire 10V sur la CG pour le rendre passant avec un canal quasiment au même potentiel que le drain (4V). Côté transistor de sélection, une tension assez basse est appliquée sur la grille de sélection (SG), environ 1V, pour créer un canal qui sera au même potentiel que la source (0V). Toute la différence de potentiel entre la source et le drain est confinée entre le point mémoire et le transistor de sélection générant ainsi des électrons énergétiques qui seront injectés dans la grille flottante grâce à la forte polarisation de la grille de contrôle (Figure 1.10). La programmation SSI se différencie de la programmation CHE par une génération d'électrons chauds proche de la source de l'empilement mémoire, d'où le nom de *Source Side Injection,* permettant d'augmenter considérablement l'efficacité d'injection en assurant qu'une grande partie des électrons du canal soit injectée dans la grille flottante. Ce mécanisme offre une rapidité de programmation similaire à la programmation par CHE tout en réduisant considérablement la consommation de courant sans pour autant atteindre les niveaux d'une programmation par effet Fowler-Nordheim. Le choix de ce mécanisme de programmation sera régi soit par l'architecture, soit par l'application visée par le produit final.

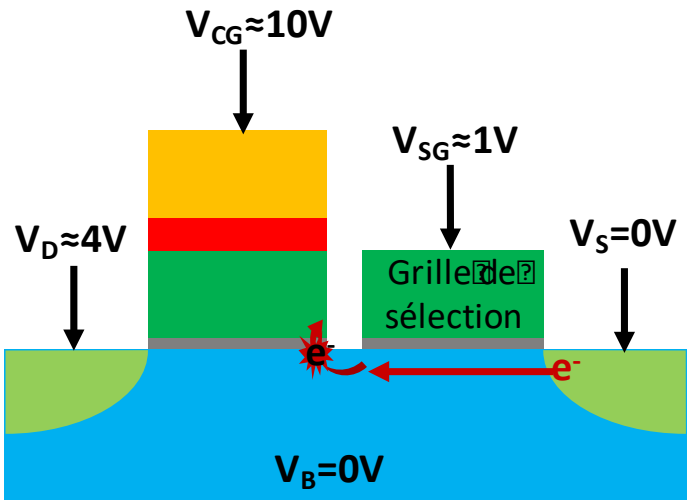


Figure 1.10 - Représentation du mécanisme de programmation par SSI.

3. Les mécanismes d'effacement

Dans la littérature, quatre façons différentes sont décrites pour effacer une cellule mémoire à grille flottante.

- **Effacement par effet Fowler-Nordheim**

L'effacement par effet Fowler-Nordheim repose sur le même principe que la programmation éponyme. L'objectif étant de chasser les électrons stockés dans la grille flottante, on applique cette fois-ci une forte tension négative (autour de -18V) sur la grille de contrôle qui va générer un champ électrique dans l'oxyde de tunnel. Ce champ électrique va permettre aux électrons piégés de passer vers le substrat. La source, le drain et le substrat sont mis à la masse (Figure 1.11). Tout comme lors de la phase de programmation, sa consommation peut être négligeable et son effacement est uniforme tout le long du canal.

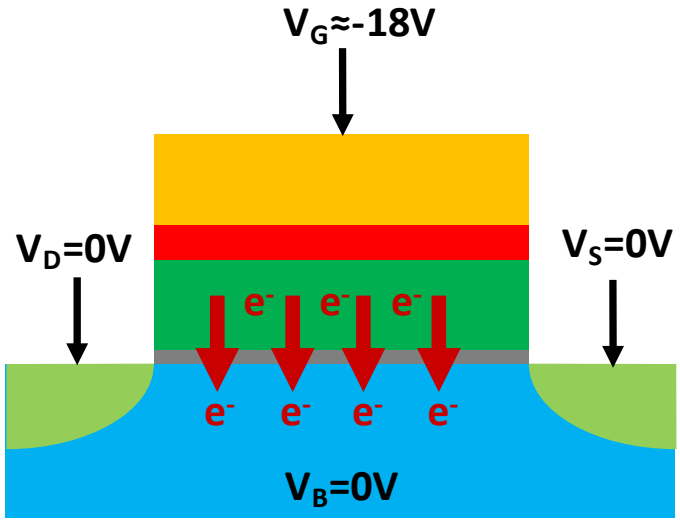


Figure 1.11 - Représentation du mécanisme d'effacement par FN.

- **Effacement par Injection de trous chauds**

L'effacement par injection de trous chauds (*Hot Holes Injection* ou HHI) se fait en polarisant en inverse la jonction substrat-drain (Figure 1.12). Ici les trous générés par ionisation par impact (ou effet tunnel bande à bande) au niveau du drain sont en partie collectés pour la grille flottante en appliquant une tension négative sur la grille de contrôle. Les trous ainsi injectés dans la grille flottante neutralisent la charge correspondant aux électrons stockés.

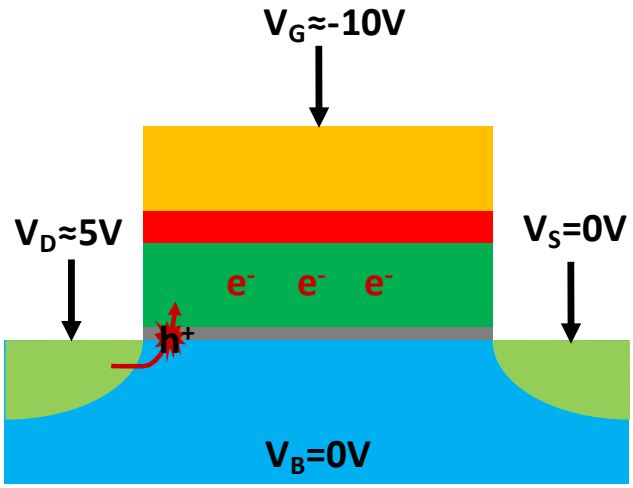


Figure 1.12 - Représentation du mécanisme d'effacement par HHI.

- **Effacement par la source**

Il est aussi possible d'effacer les cellules par la source en appliquant une tension positive élevée sur la source, les autres électrodes n'étant pas polarisées. Les électrons contenus dans la grille flottante sont alors injectés dans la source au travers de l'oxyde tunnel (Figure 1.13). Cette méthode est peu utilisée, car elle nécessite d'appliquer une forte tension sur la source, son effacement est localisé proche de la source et dépend de la zone de chevauchement entre la source et la grille flottante.

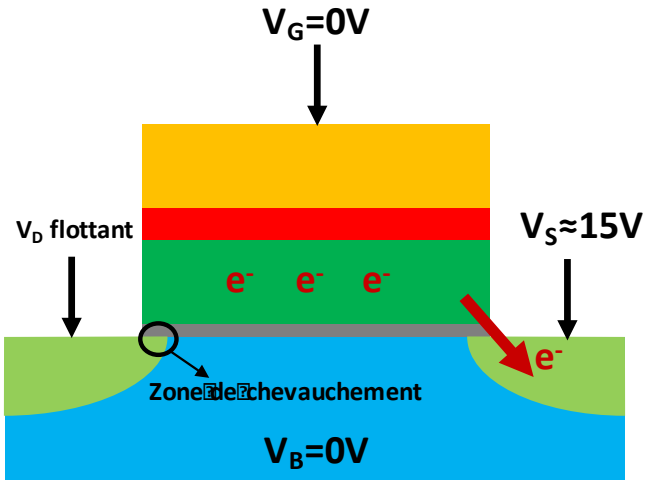


Figure 1.13 - Représentation du mécanisme d'effacement par la source.

- **Effacement par la grille et la source**

Cette méthode d'effacement est la combinaison de l'effacement par la source couplée avec une polarisation de grille. En polarisant négativement la grille (-10V) il est possible de diminuer la tension de source initialement à 15V dans le cas d'un effacement par la source à seulement 5V. Grâce à cette méthode, la différence de potentiels d'une quinzaine de volts entre la grille et la source est conservée et la différence de potentiels entre la grille et le substrat permet l'utilisation d'une plus grande zone d'effacement. Comme pour le mécanisme d'effacement par la source, le drain est gardé flottant pour éviter le passage du courant entre le drain et la source (Figure 1.14).

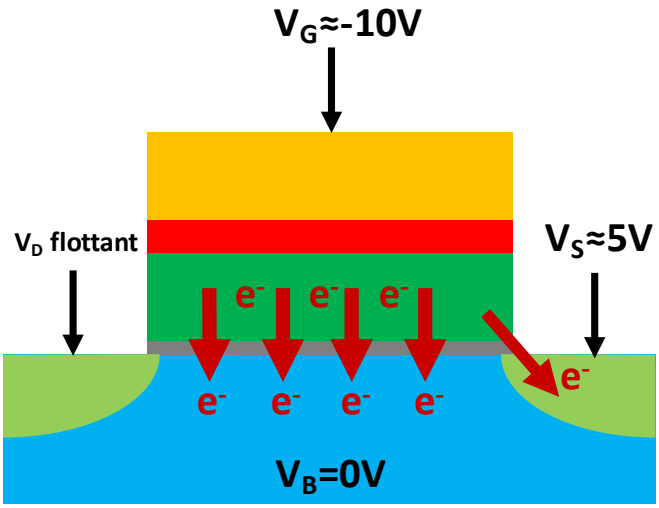


Figure 1.14 - Représentation du mécanisme d'effacement par la source et la grille.

IV. Limites de miniaturisation des mémoires Flash et perspectives envisagées

L'évolution du marché des semi-conducteurs et la prospérité de la recherche dépendent de l'essor de nouvelles applications. Depuis l'invention des mémoires Flash, l'architecture des dispositifs et leurs matériaux ont connu d'énorme progrès pour se rapprocher du dispositif "idéal". Plusieurs types de mémoires ont été inventés dans le but de maximiser certaines propriétés spécifiques. C'est pour ces raisons que le développement des mémoires a été orienté dans différentes directions [20], [21]. L'objectif de ce paragraphe est de lister les principaux obstacles à la miniaturisation des cellules Flash et d'évoquer les différents scénarios envisagés afin de permettre des évolutions ultérieures des technologies mémoires non volatiles.

1. Les principaux effets parasites dans les cellules mémoires

a) Courant de fuite induit par le stress : SILC (*Stress Induced Leackage Current*)

Durant toute la vie de la cellule mémoire EEPROM/Flash, des porteurs vont transiter à travers l'oxyde tunnel durant les phases de programmations/effacements. Les passages successifs de porteurs fortement énergétiques vont engendrer l'apparition de défauts dans et aux interfaces

de l'oxyde tunnel, diminuant lentement sa capacité à conserver les charges stockées dans la grille flottante (Figure 1.15).

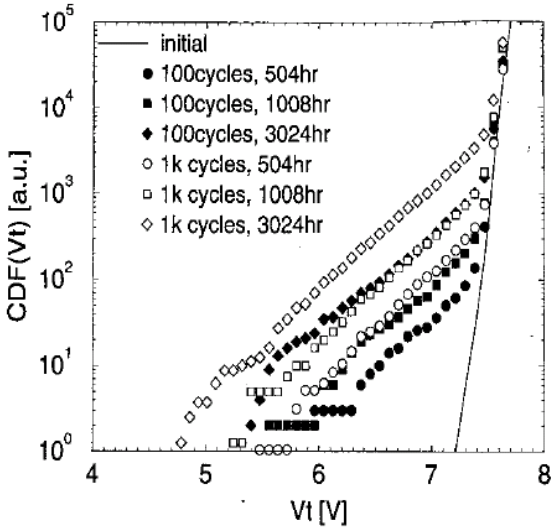


Figure 1.15 - Distribution cumulative de bits en fonction de la tension de seuil pour plusieurs conditions de cyclage [22]

Ce phénomène de SILC, *Stress Induced Leackage Current*, est d'autant plus prononcé que l'épaisseur de l'oxyde de tunnel est faible [22]. Par conséquence les fabricants n'ont d'autre alternative que de limiter la réduction d'épaisseur de l'oxyde tunnel à 6-7 nm, ce qui affecte la miniaturisation des composants.

b) Les effets liés aux canaux courts : SCE et DIBL

Avec la miniaturisation des composants, des effets liés à l'utilisation des canaux courts apparaissent [23]–[26]. Deux principaux phénomènes sont présents aussi bien dans les transistors que dans les empilements à grille flottante :

- L'effet canal court ou SCE (*Short Channel Effect*) : il apparaît lorsque la longueur L du transistor devient trop faible. Ainsi les zones de charge d'espace de la source et du drain se rejoignent créant un canal de façon involontaire.
- DIBL (*Drain Induced Barrier Leackage*) : il apparaît lorsqu'une tension est appliquée sur le drain. Ce qui entraîne une diminution de la tension de seuil et la dégradation de la pente sous-seuil. En raison de DIBL, le courant de fuite augmente et la consommation d'énergie atteint des valeurs incompatibles avec les exigences des nœuds technologiques avancés.

Ces deux phénomènes se traduisent par une perte de contrôle du canal, une forte augmentation de la pente sous le seuil et d'importantes variations de la tension de seuil (V_T) en fonction des tensions de drain.

c) Les effets liés aux couplages électrostatiques

Comme cela a déjà été mentionné, la cellule Flash repose sur le couplage électrostatique entre la grille flottante et le canal du transistor. Par conséquent, tout effet parasite venant moduler le potentiel de la grille flottante a un effet néfaste sur le fonctionnement des cellules mémoires.

Comme le montre la Figure 1.16, l'un des moyens retenus pour assurer un bon couplage avec la grille flottante consiste à enrober cette dernière avec la grille de contrôle. Cela permet ainsi, d'étendre la surface de couplage électrostatique aux flancs de la structure. Cependant, en-deçà du nœud technologique 35 nm, la réduction d'échelle provoque une diminution de l'espacement entre les cellules, ne permettant plus le recouvrement latéral de la grille flottante par la grille de contrôle.

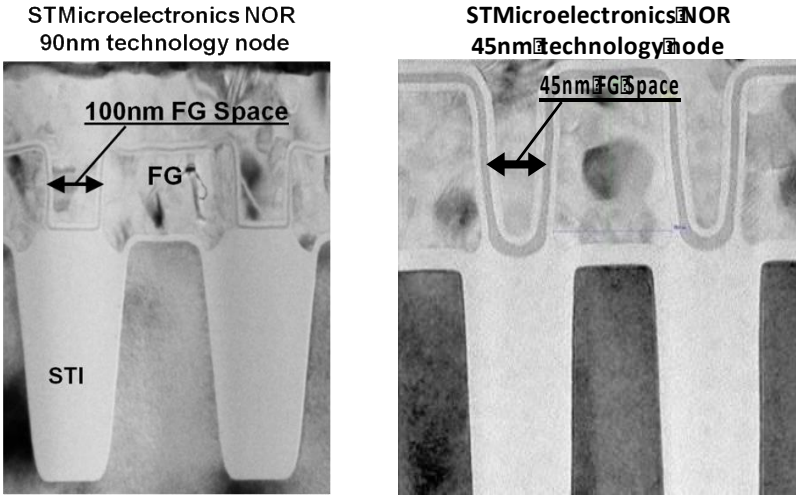


Figure 1.16 - Photos TEM de la mémoire Flash 90nm NOR de STMicroelectronics (à droite) et la Flash NOR sub-45nm de STMicroelectronics (à gauche) [Interne ST].

2. Solutions envisagées : évolutions et ruptures technologiques

Compte tenu des points bloquants qui viennent d'être mentionnés, différentes architectures mémoires sont apparues. Ces cellules mémoires privilégient d'autres modes de stockage de l'information. Parmi les différentes variétés des technologies envisagées, on retrouve :

- Les mémoires FRAM (*Ferroelectric Random Access Memory*) : Dans les années 50, le monde de la recherche commence à s'intéresser aux matériaux ferroélectriques et leur possibilité d'être utilisés dans des mémoires [27]. L'architecture d'une mémoire FRAM, similaire à celle des mémoires DRAM présente la particularité d'intégrer une capacité ferroélectrique comme élément de stockage (Figure 1.17.a et 1.17.b) [28]. Les matériaux ferroélectriques possèdent un cycle d'hystérésis de polarisation en fonction du champ électrique (Figure 1.17.c) apportant un comportement non-linéaire.

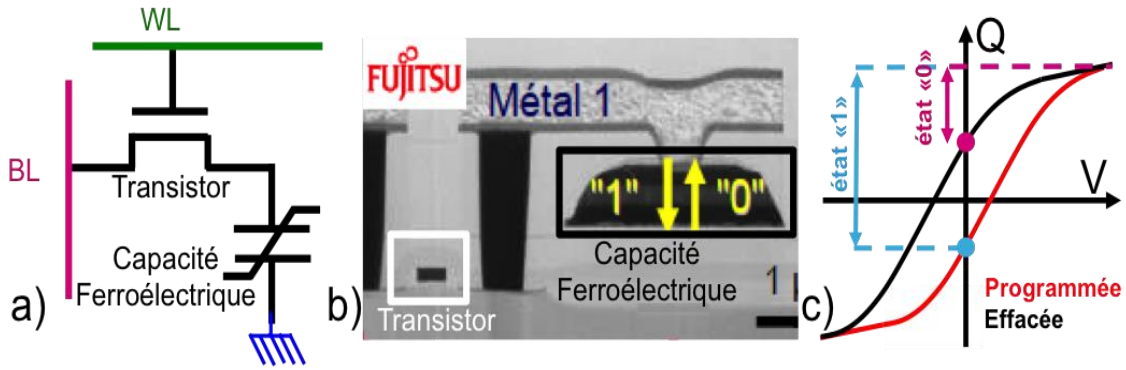


Figure 1.17 - a) Schéma électrique b) Photo TEM (Fijitsu) c) Cycle d'hystérésis d'un oxyde ferroélectrique.

- Les mémoires à changement de phase ou PCM (*Phase Change Memory*) : En 1969, A.V. Pohn met en avant la possibilité d'utiliser un alliage chalcogène en guise de mémoire [29]. Ces alliages présentent des caractéristiques physiques (indice optique, conductivité) différentes en fonction de leur phase cristalline. Ils peuvent passer d'un état solide amorphe (isolant) à l'état cristallin (conducteur) en fonction des impulsions électriques auxquelles ils sont soumis. L'alliage chalcogène (exemple : le GST pour Germanium (Ge) / Antimoine (Sb) / Tellure (Te)) est placé entre une électrode supérieure et un élément chauffant (radiateur), relié à une électrode inférieure. À l'état initial, le GST est dans l'état cristallin. Pour programmer la cellule, il est nécessaire de forcer un courant entre les deux électrodes, faisant chauffer le radiateur (environ 400°C) amenant le chalcogène à un état cristallin. Ainsi, lors de la phase de lecture qui suit, le dispositif conducteur laissera plus facilement passer un courant. Pour l'effacement, le passage d'un courant plus élevé afin que la température du radiateur atteigne 600°C environ permet à la couche de GST de passer à l'état liquide puis à l'état solide amorphe grâce à un refroidissement brutal. La Figure 1.18 montre, à l'aide d'un schéma et d'une photo TEM, la différence entre l'état programmé (cristallin) et l'état effacé (amorphe). Depuis les années 2000, de nombreuse entreprises (Intel, Samsung, NXP, TSMC, STMicroelectronics …) ont commencé à investir massivement dans la recherche en vue de la commercialisation de ce type de mémoire [30]–[33].

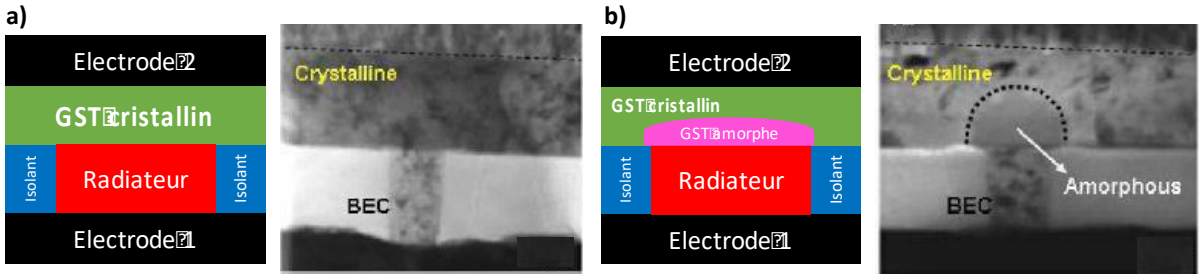


Figure 1.18 - Schéma et photo TEM d'une mémoire PCM - a) dans l'état amorphe ("0") et b) dans l'état cristallin ("1"). [32]

- Les mémoires MRAM (Magnetic RAM) : Ces mémoires sont basées sur un stockage par orientation magnétique. Le changement d'état se fait par la

commutation du spin des électrons. Le point mémoire se compose d'un oxyde tunnel situé entre deux matériaux ferromagnétiques et deux électrodes (Figure 1.19) [34], [35]. Un des deux matériaux ferromagnétiques sert de référence : l'orientation de son aimantation reste constante (en rouge sur les Figure 1.19.a et b). L'autre (en vert) verra le spin de ses électrons s'inverser lorsqu'un champ magnétique sera appliqué, par les lignes d'écriture WL, entre les deux couches ferromagnétiques (Figure 1.19b) ; c'est l'écriture ou l'effacement selon l'état initial. La lecture se fait par une simple mesure de résistance et donne deux valeurs bien distinctes suivant l'orientation de l'aimantation par rapport à une référence.

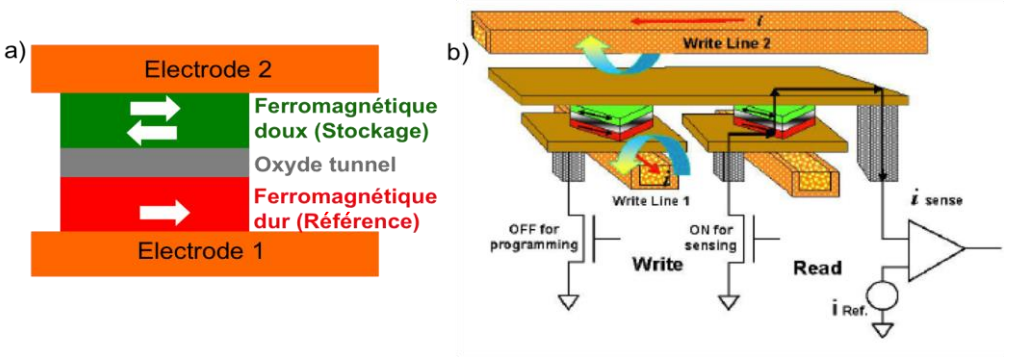


Figure 1.19 - a) Composition b) Fonctionnement d'une mémoire MRAM [36]**.**

- Les mémoires OxRAM : Basées sur structure MIM (Métal Isolent Métal), elles se composent de deux électrodes métalliques et d'un oxyde de métal de transition comme le NiO [37], le HfO_2 ou TiO_2 [38]. Un changement de résistivité d'un état fortement résistif (appelé HRS) à un état faiblement résistif (LRS) a été démontré pour la première fois dans les années 60 [39]. En utilisation standard, le changement d'état de HRS vers LRS est appelé « l'étape de set » et s'effectue par l'application d'une tension au moins supérieure à la tension de commutation notée V_{SET}. Le passage de LRS vers HRS, appelé « l'étape de reset », est observé lorsque la tension V_{RESET} est atteinte. Les différents cycles de programmation sont détaillés dans la Figure 1.20. Toutefois, la première commutation de l'état vierge à l'état LRS nécessite une tension supérieure à la moyenne. Cette étape « d'électro *forming* » est une étape clé dans la vie de la cellule. Elle est actuellement l'un des freins aux déploiements industrielles de cette technologie.

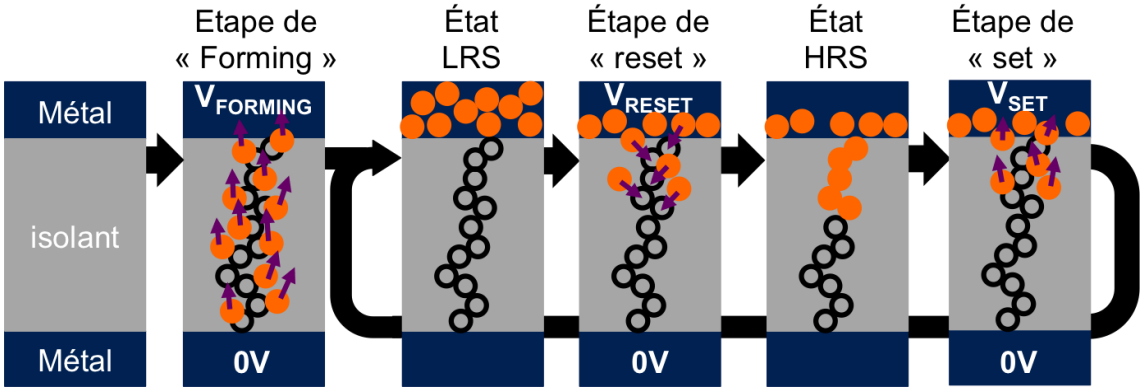


Figure 1.20 - Fonctionnement d'une mémoire OxRAM

3. Développement d'architecture 3D

L'arrivée aux derniers nœuds technologiques (i.e. 8nm et en deçà) ne signifie pas la fin du développement technologique des mémoires non volatiles. En effet, le développement d'architectures tridimensionnelles devrait permettre de poursuivre l'augmentation de la densité d'intégration, notamment pour les architectures de type NAND. Actuellement, les principaux types de mémoire 3D sont des architectures de type "gate-all-around" à canal vertical (Figure 1.21a) ou à grille verticale (Figure 1.21b). Ces architectures permettent l'empilement de plusieurs plans mémoires les uns sur les autres afin de réduire de façon conséquente l'encombrement des cellules mémoires individuelles. Notons que pour ces deux types d'intégration la technologie des mémoires à piégeage de charges est préconisée. Cependant, une rupture technologique est envisagée par le remplacement de la technologie par piégeage de charges par la technologie ReRAM ou PCM.

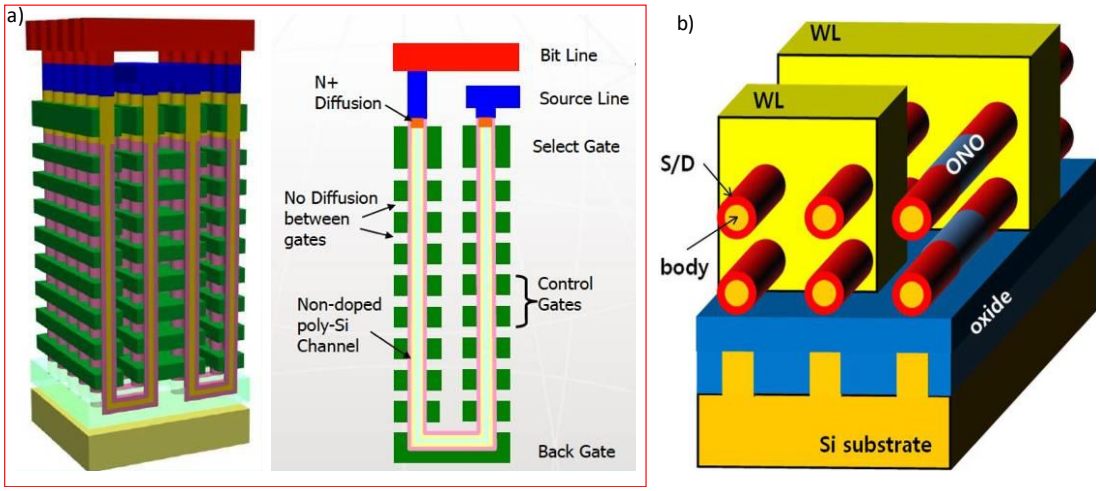


Figure 1.21 : Schéma d'une architecture 3D avec a) canal vertical b) grille verticale [40]

V. Les mémoires à grille flottante dédiées aux applications basse consommation

Les mémoires Flash et EEPROM sont les mémoires non volatiles les plus utilisées dans les appareils électroniques qui nous entourent. Cependant, de nouvelles mémoires font leur apparition avec l'arrivée de nombreux dispositifs nomades. Pour cibler ce marché, les mémoires

doivent présenter une consommation électrique en baisse afin de préserver le plus possible leur niveau de batterie.

Comme indiqué au début de ce chapitre, la mémoire Flash NOR est programmée avec le mécanisme d'injection d'électrons chauds qui présente une forte consommation de courant. Pour diminuer la consommation durant la programmation, différentes méthodes existent comme optimiser les tensions appliquées ou encore modifier l'architecture de la mémoire. La première solution est généralement insuffisante, car une diminution de la consommation affectera les performances (fenêtre de programmation, endurance, rétention) de la cellule mémoire. Il est préférable de travailler sur de nouvelles architectures mémoires permettant une meilleure injection de charges et une diminution de la consommation tout en maintenant la fenêtre de programmation souhaitée. Ainsi deux architectures mémoires non volatiles avec des consommations plus faibles que la Flash standard sont particulièrement répandues dans la littérature : les mémoires à deux transistors (2T) et la *Split Gate*.

1. La mémoire à deux transistors (2T)

La première cellule est la mémoire à deux transistors appelée 2T (fabriquée entre autres par NXP [41], Samsung [42], Philips [43] ou Actel [44]). Cette cellule est composée d'un transistor de sélection permettant d'accéder à un second transistor qui sert de point mémoire, tel que présenté en Figure 1.22. Les avantages de cette architecture sont sa bonne granularité et sa faible consommation grâce à la programmation par mécanisme Fowler Nordheim.

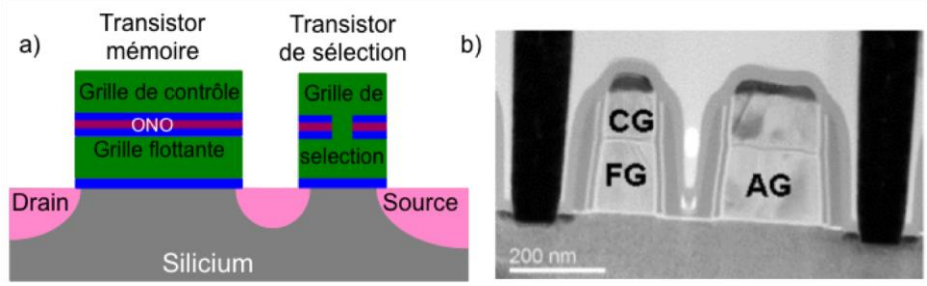


Figure 1.22 - a) Schéma b) Coupe TEM de l'architecture 2T.

Lors de la lecture, une faible tension est appliquée sur le drain, la source est mise à la masse et le canal du transistor de sélection est inversé grâce à une polarisation de 3V sur sa grille. Une tension particulière est appliquée sur la grille du transistor mémoire qui va permettre de laisser passer du courant si la cellule est effacée ou de bloquer le transistor si elle est programmée. La présence du transistor de sélection permet d'accéder à un seul point mémoire. L'effacement de la cellule se fait par Fowler Nordheim. Son architecture lui donne la possibilité d'être programmée par FN d'où la faible consommation en programmation.

La consommation très faible de la mémoire 2T reste son atout majeur, mais le temps de programmation et les tensions utilisées sont élevées. Néanmoins, l'inconvénient majeur réside dans sa taille due à son architecture à deux transistors.

2. La mémoire Split Gate

L'architecture Split Gate est apparue dans les années 90 pour les applications basse consommation [45]. L'architecture de cette mémoire a ensuite évolué afin d'améliorer ses

performances. La Figure 1.23 résume l'architecture de trois générations de Split Gate [46]. La première et deuxième génération de Split Gate, sont composées d'une grille flottante ainsi que d'une grille de sélection (SG) servant aussi de grille de contrôle. La lecture est identique à une Flash, le drain est polarisé à 1V, la source à 0V pour permettre au courant de traverser. Une tension particulière est appliquée sur la SG. L'effacement se fait par Fowler Nordheim à l'aide d'une forte tension positive appliquée sur la SG. Grâce à cette architecture une forte tension négative n'est plus indispensable pour effacer, ce qui permet de simplifier la périphérie nécessaire à la création des fortes tensions d'écriture/effacement. La méthode de programmation utilisée est la SSI (paragraphe III.2). Cette méthode est typique à cette architecture, car elle est possible uniquement grâce à la présence d'un transistor de sélection. La troisième génération de la Split Gate introduit une nouvelle grille dans son l'architecture : l'erase gate (EG), ainsi que l'utilisation d'une vraie grille de contrôle (CG). L'effacement de cette structure sera fait via cette nouvelle grille EG, simplifiant la circuiterie qui adresse la zone mémoire. La lecture et la programmation restent similaires aux générations précédentes.

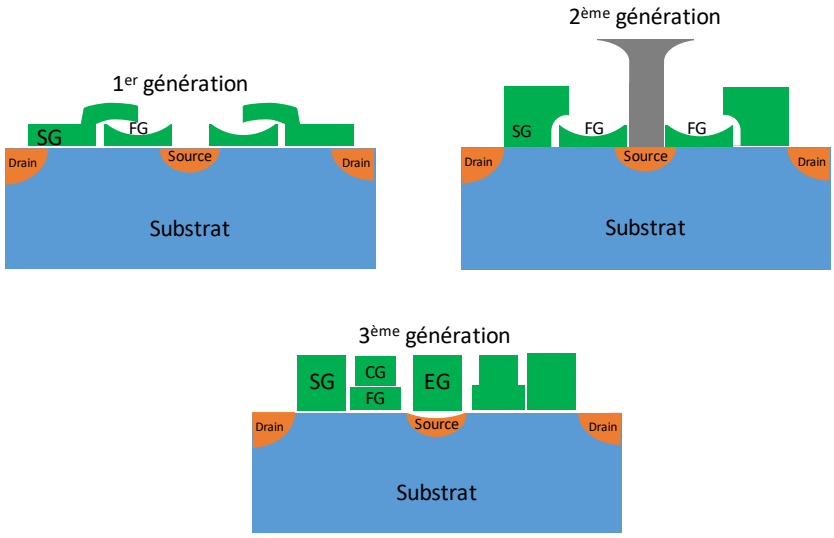


Figure 1.23 - Architecture des trois générations de Split Gate.

La consommation très faible et la densité de la mémoire Split Gate (3ème génération) reste son atout majeur. Le prix à payer, est un procédé de fabrication complexe qui rend le coût de fabrication très élevé.

Cependant les différentes solutions proposées auparavant sont protégées par des brevets et la fabrication de ces architectures par STMicroelectronics-Rousset entraînera un paiement de redevance. Ces redevances amplifieront le coût de fabrication ce qui limitera la rentabilité de la production. C'est pour cette raison qu'une nouvelle solution "eSTM", appartenant à STMicroelectronics, a été proposée et dont la description sera faite dans la prochaine partie.

3. Nouvelle architecture mémoire ultra basse consommation : eSTM

Pour adresser les applications basse consommation, les ingénieurs de STMicroelectronics de Rousset ont développé une nouvelle architecture mémoire en se basant sur la cellule mémoire 2T (paragraphe V.1). Cette architecture mémoire a pour objectifs :

- Intégrer une tranchée en polysilicium comme transistor de sélection.

- Avoir un procédé de fabrication assez simple qui se rapproche de celui d'une cellule Flash standard.
- Garder la même densité que la mémoire Flash.

Cette partie décrit, dans un premier temps, l'état de l'art des tranchées en polysilicium. Ensuite, nous présenterons la nouvelle architecture de la cellule eSTM.

a) L'état de l'art des tranchées en polysilicium

i) Tranchées utilisées comme isolant

L'isolation des structures voisines les unes par rapport aux autres est de plus en plus importante avec la diminution des nœuds technologiques. Des structures utilisant des tranchées en polysilicium ont été proposées pour diminuer les différents effets de couplages parasites. La Figure 1.24 montre deux applications différentes d'isolation. La première application [47] (Figure 1.24a) consiste à isoler deux transistors mémoires en ajoutant, sous le STI en SiO_2, une tranchée polysilicium d'une taille de 1.5 à 3µm de profondeur et une largeur de 160 à 180nm. Dans le deuxième exemple (Figure 1.24b) [48], une tranchée est utilisée comme anneau isolant pour encercler une structure. La profondeur de cette tranchée varie entre 1 et 6µm mais aucun détail sur la largeur n'est donné. Dans les deux cas, la présence d'une tranchée de polysilicium améliore les performances des dispositifs.

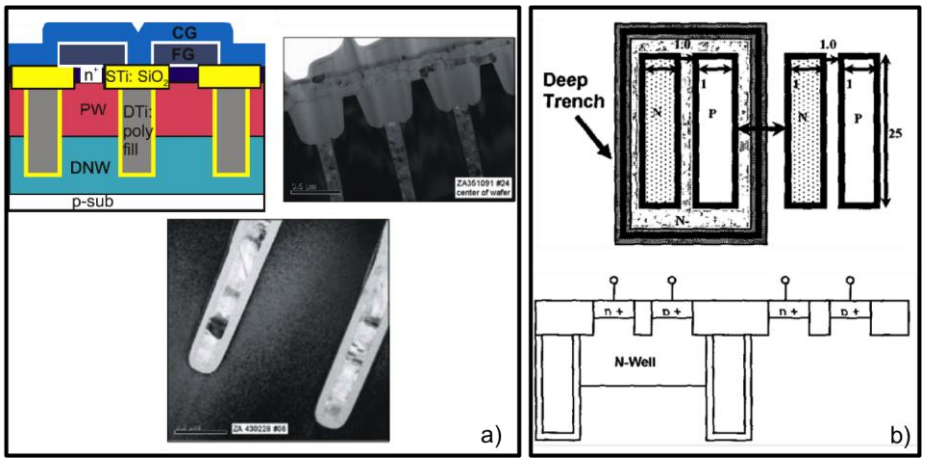


Figure 1.24 - Deux utilisations de tranchée en polysilicium pour diminuer les couplages parasitiques [45] [46]

ii) Tranchées utilisées comme inductance

Les tranchées de polysilicium ont aussi été utilisées afin d'améliorer les performances des inductances [49]. Les schémas de la Figure 1.25 montrent la succession d'étapes permettant la réalisation d'une inductance intégrant des tranchées d'une profondeur de 6µm et d'une largeur de 2µm. La présence de ces tranchées en polysilicium permet une augmentation de :

- La fréquence de résonnance grâce à la diminution de la capacité de couplage
- Du facteur de qualité en diminuant les pertes résistives

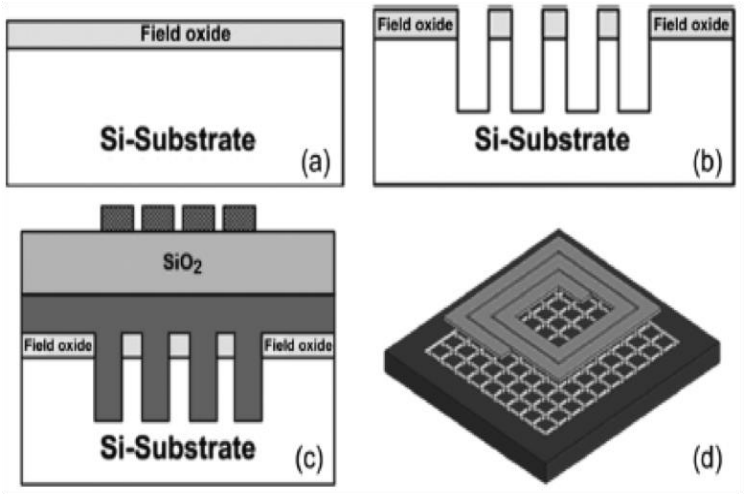


Figure 1.25 - Etapes de réalisation des tranchées pour l'amélioration d'une inductance.

a) Substrat de départ, b) Gravure des tranchées, c) Remplissage des tranchées, d) Inductance intégrant des tranchées. [47]

iii) Tranchées utilisées comme capacité pour les mémoires DRAM

L'une des applications les plus spectaculaires des tranchées de polysilicium sont les capacités des mémoires volatiles de type DRAM [50]. Dans ces mémoires composées d'un transistor et d'une capacité, beaucoup d'évolutions ont été apportées afin d'améliorer les performances et/ou la taille de la cellule. L'une des évolutions remarquables est l'utilisation de capacité enterrée comme le décrit la Figure 1.26.

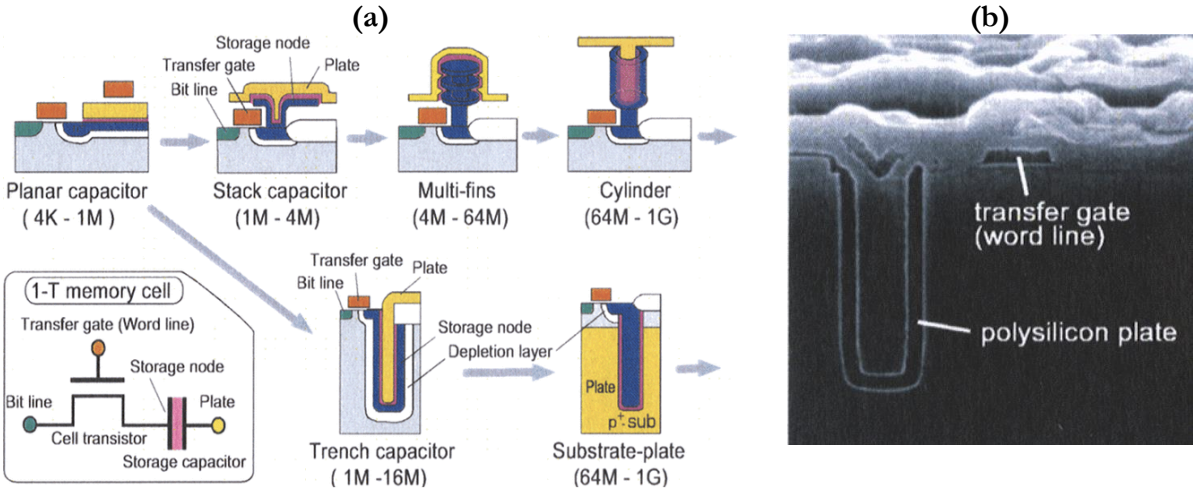


Figure 1.26 - a) Evolution des DRAM b) Coupe TEM d'une réalisation de tranchée de polysilicium utilisée comme capacité dans une DRAM [50]

iv) Tranchées utilisées comme transistor à grille flottante

La tranchée en polysilicium a également déjà été utilisée comme grille flottante de structures mémoires (Figure 1.27) [51]–[53]. Cette mémoire est une architecture de type Split Gate. La tranchée de profondeur d'environ 200nm et d'une largeur de 280nm (Figure 1.27a), est remplie de polysilicium déposé après la croissance de l'oxyde (oxyde tunnel). Le polysilicium est

enfin gravé afin de n'en laisser que le long des parois verticales de la tranchée, définissant les grilles flottantes des cellules mémoires. Un second oxyde épais est réalisé avant un second remplissage de polysilicium, servant de contact de source au fond de la tranchée. Cette cellule se programme par SSI et s'efface par FN par la grille de contrôle (Figure 1.27b).

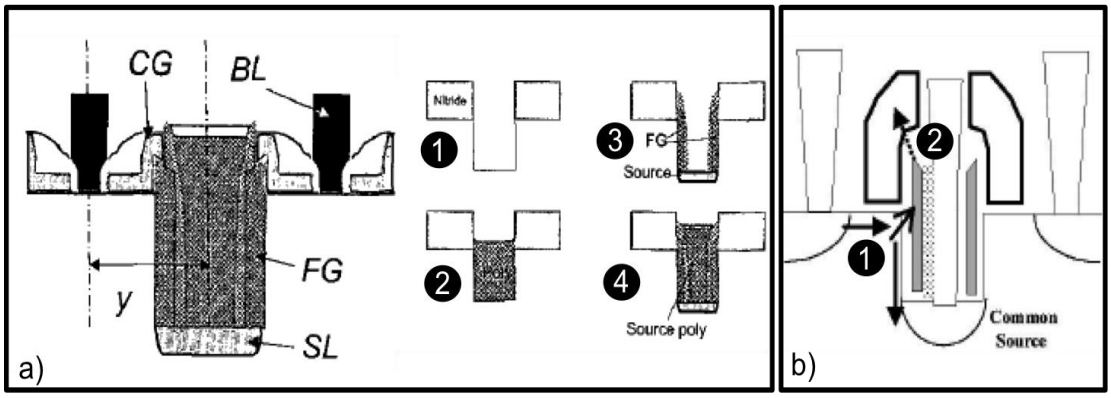
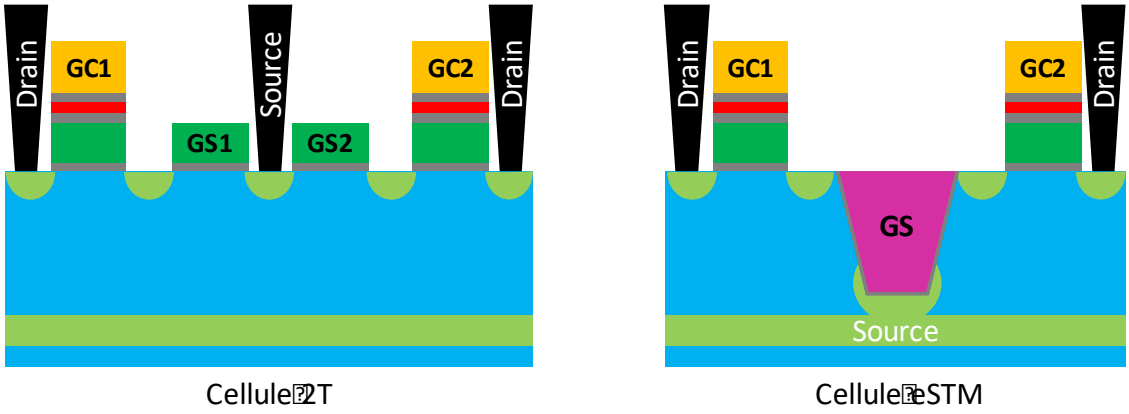


Figure 1.27 - Cellule à grille flottante enterrée - a) Procédé de fabrication b) Mécanismes de fonctionnement (programmation en 1 et effacement en 2) [51]

b) La cellule eSTM (Embedded Select Trench Memory)

Mise à part sa taille, la cellule 2T est une cellule mémoire très intéressante en termes de consommation et d'isolation par rapport à la source grâce à son transistor de sélection. De plus elle offre un procédé de fabrication compatible avec celui des mémoires Flash embarquées. En fusionnant les deux transistors en un seul transistor vertical, la cellule 2T trop volumineuse peut être transformée en une nouvelle cellule ayant la même taille qu'une cellule Flash standard. Cependant le procédé de fabrication devient plus complexe tout en restant compatible avec un procédé CMOS type Flash (Figure 1.28). Cette nouvelle cellule eSTM (*Embedded Select Trench Memory*) sera étudiée tout au long de ce manuscrit avec une description détaillée de son procédé de fabrication et les résultats électriques de cette dernière.



Cellule 2T Cellule eSTM

Figure 1.28 - Passage d'une 2T à la cellule eSTM

VI. Conclusion

Dans ce chapitre, les principales mémoires non volatiles ont été présentées. L'étude de l'état de l'art montre que les mémoires non volatiles constituent un marché stratégique pour l'industrie des semi-conducteurs. Le succès des produits de l'internet des objets ces deux dernières années accentue l'importance des dispositifs à mémoire embarquée de type Flash. Malgré les performances électriques et la fiabilité qui font des mémoires Flash la technologie de référence en matière de mémoire non volatile, ces dispositifs atteignent plusieurs limites d'intégration. Celles-ci concernent les effets de couplages parasites liés à la miniaturisation ainsi que la limite de résolution des équipements de photolithographie. Afin de poursuivre l'évolution des mémoires non volatiles, plusieurs solutions sont envisagées.

À long terme les mémoires à stockage de charges souffriront des effets parasites pour les dimensions sub-28nm (paragraphe IV.1.c). Ces problématiques vont pousser les industriels à utiliser les mémoires émergentes, notamment les mémoires à changement de phase (PCM) ou les mémoires à commutation de résistance (ReRAM). Notons que certains industriels proposent dès à présent, dans de petits volumes de production, des produits basés sur des mémoires PCM, ReRAM ou MRAM.

Sur le court ou le moyen terme, les mémoires Flash restent une technologie fiable, rentable, et performante du point de vue industriel. Dans ce sens, plusieurs innovations technologiques vont voir le jour pour préserver cette technologie aussi longtemps que possible. C'est dans cette optique que s'inscrit le développement de la nouvelle architecture eSTM.

Le sujet de cette thèse consiste d'une part à étudier l'impact de la variabilité du procédé de fabrication sur les performances électriques de la nouvelle architecture eSTM. D'autre part, à proposer des solutions pour optimiser la variabilité du procédé de fabrication, et de proposer des boucles de régulations pour un meilleur contrôle en ligne. Et finalement une campagne de test électrique sera effectuée pour évaluer l'impact des optimisations effectuées sur la mémoire eSTM.

Chapitre 2 : ETUDE DE LA CELLULE eSTM – EMBEDDED SELECT TRENCH MEMORY

Sommaire

L'objectif principal de ce chapitre est d'expliquer le fonctionnement de la cellule eSTM. Avant de décrire son fonctionnement électrique, nous allons faire une description de son procédé de fabrication.

I. Procédé de fabrication de l'eSTM

Les procédés de fabrication de la microélectronique peuvent être divisés en deux parties : FEOL (Front End Of Line) qui regroupe toutes les étapes de la réalisation des cellules élémentaires et la partie BEOL (Back End Of Line) qui va réaliser les contacts et les différents niveaux de lignes de métaux afin de relier la mémoire au monde extérieur. Dans cette partie, nous allons nous concentrer uniquement sur la partie FEOL.

La réalisation de l'eSTM se fait grâce à une succession d'étapes, proches de celles de la cellule Flash, mais avec des étapes supplémentaires pour intégrer la tranchée en polysilicium. Il convient aussi de modifier certaines briques existantes comme la taille des points mémoires ou des contacts. La Figure 2.1 met en avant les étapes supplémentaires et celles à modifier par rapport au procédé de fabrication de la Flash NOR.

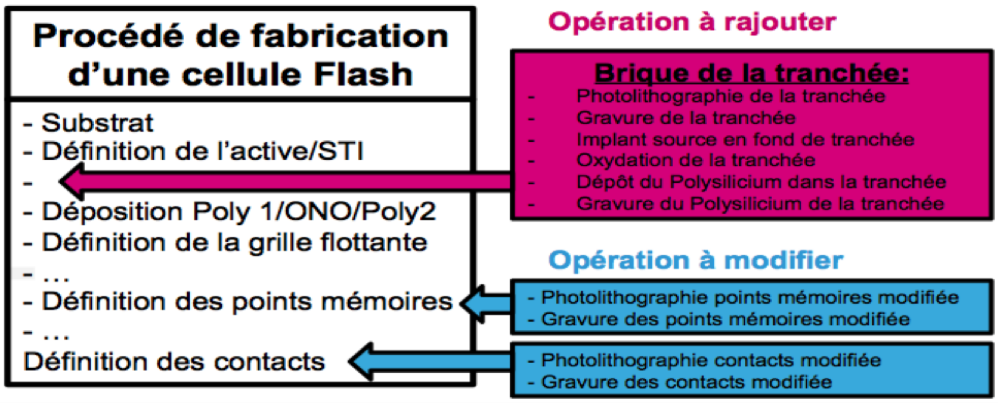


Figure 2.1 - Flux de procédé de fabrication simplifié d'une Flash et modifications à apporter pour la réalisation d'une eSTM

Le processus de fabrication débute avec la définition des zones d'active qui sont isolées par les tranchées d'isolation STI (Shallow Trench Isolation). Pour cela, une couche d'oxyde sacrificiel est obtenue par croissance sur un substrat de silicium. Une couche de nitrure, servant de couche de protection est déposée au-dessus de l'oxyde sacrificiel (Figure 2.2a).

Après un dépôt de résine photosensible (Figure 2.2b), un réticule est utilisé afin de permettre à l'étape de photolithographie de protéger la résine dans les futures zones d'active et ainsi dévoiler les zones STI à graver (Figure 2.2c). Après cette étape de photolithographie, l'empilement substrat / oxyde sacrificiel / nitrure / résine est gravé laissant apparaître des tranchées dans le silicium (Figure 2.2d). L'utilisation du nitrure permet d'éviter la gravure de l'empilement après consommation de la résine dans les zones d'active.

Figure 2.2 - a) Dépôt des couches de protection b) Dépôt de la résine c) Définition des zones d'active d) Gravure des zones STI.

Ces tranchées de silicium sont ensuite remplies par un oxyde, en effectuant une oxydation thermique suivie d'un dépôt d'oxyde HDP (High Density Plasma) afin de garantir un remplissage uniforme, sans vide à l'intérieur (Figure 2.3a). Le rôle des tranchées STI est d'isoler les transistors les uns par rapport aux autres. Afin de retirer une partie du surplus d'oxyde, une étape de CMP (Chemical Mechanical Polishing) est nécessaire. Son rôle consiste à retirer uniformément l'oxyde en se servant du nitrure comme couche d'arrêt puis une gravure humide est utilisée pour retirer le nitrure, ramenant ainsi le niveau du STI au même niveau que celui de l'oxyde sacrificiel (Figure 2.3b).

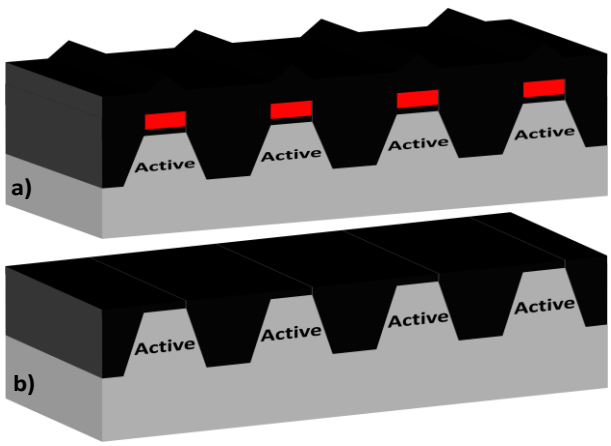


Figure 2.3 - a) Remplissage des tranchées STI par de l'oxyde b) Retrait du nitrure et du surplus par CMP puis gravure humide.

Après avoir isolés les transistors entre eux grâce au STI, un implant dopé N (appelé Niso ou triple well) est réalisé en profondeur. L'objectif est d'isoler les transistors par rapport au substrat (Figure 2.4a). Un second implant de type P est effectué dans la zone mémoire afin de doper les caissons des futurs transistors pour leur bon fonctionnement (Figure 2.4b).

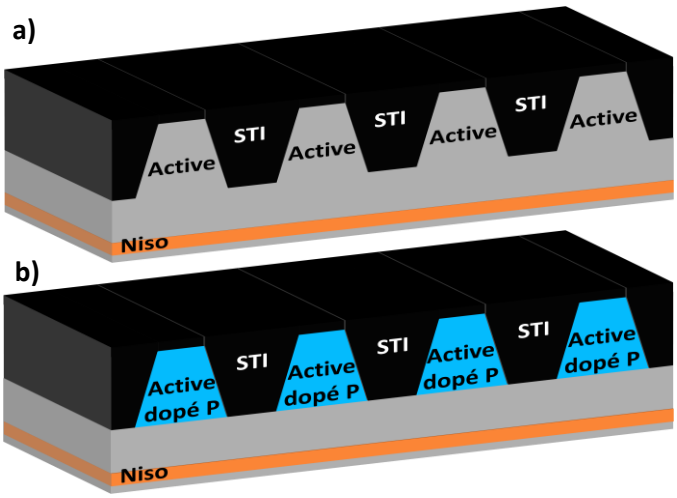


Figure 2.4 - a) Définition de l'implant isolant de Type N (Niso) b) Définition de l'implant des caissons de type P.

Après la définition de la zone Active/STI, nous devons creuser une tranchée perpendiculaire aux lignes d'active. Pour cela nous avons besoin d'un masque de protection qui va permettre de protéger les zones que l'on ne souhaite pas graver. Ce masque de protection est composé (Figure 2.5) :

- D'une couche AHM (Ashable Hard Mask) comparable à de la résine
- D'une couche de NFARL (Nitride Free Anti Reflective Layer)
- D'une couche de résine 193nm

L'utilisation du NFARL a pour objectif d'éviter la réflexion du rayonnement laser, pendant la photolithographie, sur le AHM et ainsi ne pas impacter la définition des tranchées. La couche de carbone AHM sert à protéger les zones qui ne doivent pas être gravées. De plus elle est simple à retirer une fois l'étape de gravure terminée.

En production l'utilisation du NFARL est accompagnée par une couche du BARC (Bottom Anti Reflective Coating) entre la résine et le NFARL. Cette couche anti reflet assure l'adhérence avec la résine. Dans notre cas, afin de s'affranchir d'une couche en plus à graver, le BARC a été remplacé par une fine pellicule de promoteur d'adhérence (HMDS : Hexamethyldisilazane).

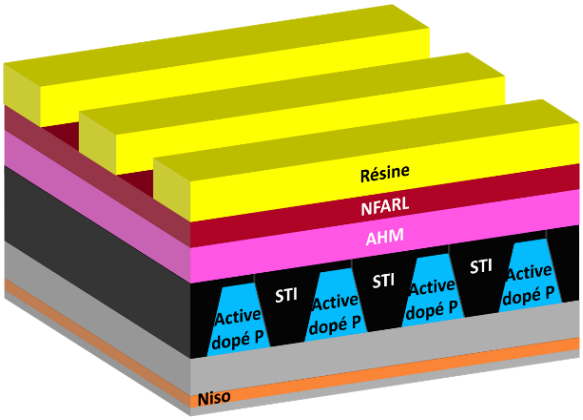


Figure 2.5 - Détail des couches de protection avant la gravure de la tranchée

La définition de la tranchée est ensuite effectuée grâce à une gravure plasma, la Figure 2.6a montre schématiquement le résultat obtenu. Nous pouvons remarquer que la profondeur de la tranchée n'est pas identique entre la coupe le long de l'active et le long du STI (Figure 2.6b). Cette différence de profondeur est due à la vitesse de gravure de l'oxyde (zone du STI) qui est plus importante que celle du silicium (zone d'active).

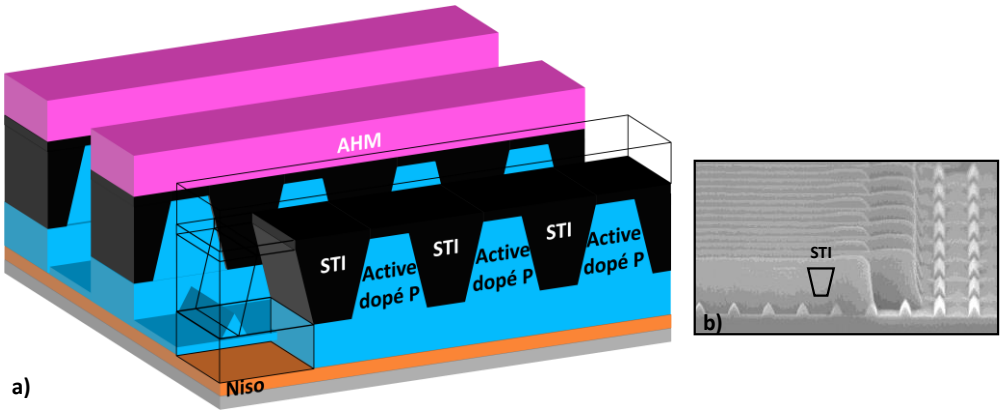


Figure 2.6 – a) Gravure plasma de la tranchée b) Coupe SEM le long de l'active et le long du STI [Interne ST]

Après la réalisation de l'étape de gravure, l'implant de la source du futur transistor vertical doit être réalisé au fond de la tranchée (Figure 2.7a). Les zones non gravées sont protégées grâce à la couche AHM restante après l'étape de gravure. Cet implant est effectué en deux fois en y ajoutant un angle d'implantation de 7° pour s'assurer que la pointe au fond de la tranchée soit totalement dopée (Figure 2.7b).

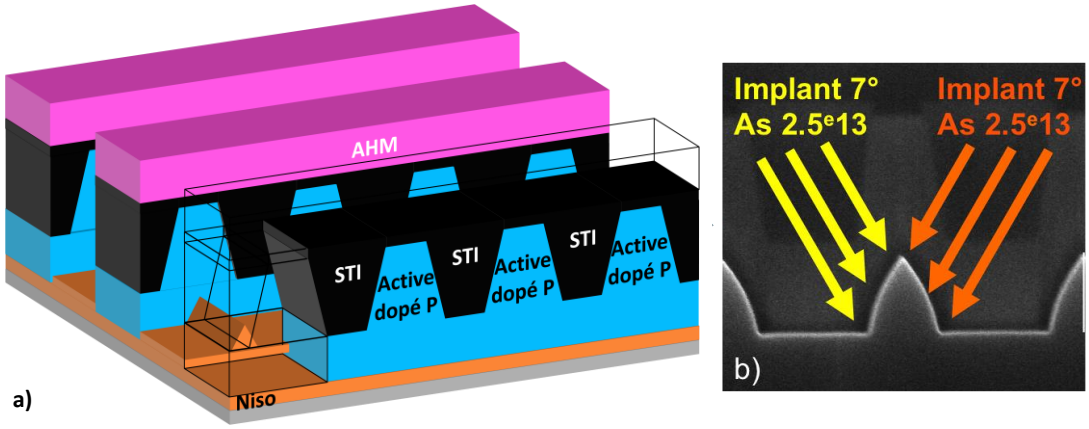


Figure 2.7 – a) Implant source du transistor vertical b) Optimisation de l'implant source

L'implant source étant réalisé, il faut à ce stade retirer le masque de protection (AHM) pour faire croître l'oxyde de grille du transistor de sélection (Figure 2.8a). La réalisation de cet oxyde est effectuée avec la méthode ISSG (In Situ Steam Generated) [54], car il doit être assez épais et uniforme tout le long de la tranchée pour éviter toutes sortes de fuites entre le transistor de sélection et le substrat.

Après l'oxydation de la tranchée l'étape suivante consiste à la remplir de polysilicium (Figure 2.8b), qui servira de grille pour le transistor de sélection. Le polysilicium déposé est dans un état amorphe et dopé In-Situ dans l'équipement de déposition, ce qui permet un dopage uniforme et évite une étape de dopage pleine plaque. Le polysilicium cristallin n'est pas utilisable ici de par sa vitesse de déposition (10 fois plus rapide que l'amorphe) qui serait propice à la création de cavités. De plus, les différents recuits présents dans la suite du processus de fabrication permettront au polysilicium de retrouver sa forme cristalline ce qui assurera une meilleur reproductibilité d'un transistor à l'autre [55].

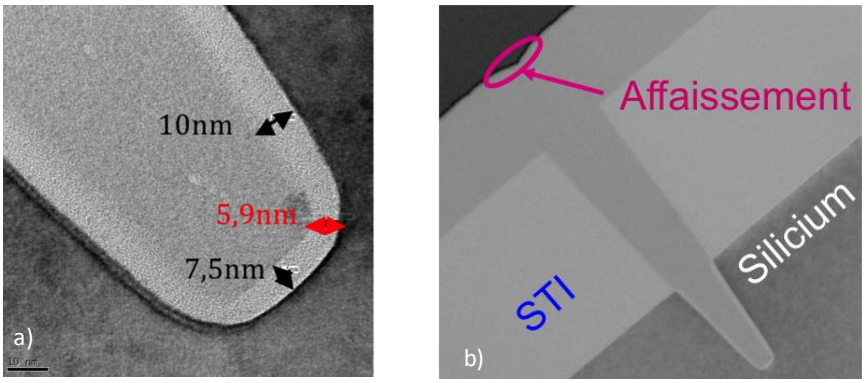


Figure 2.8 – a) Résultat de l'oxydation ISSG de la tranchée b) Coupe TEM après dépôt du Polysilicium dans la tranchée [Interne ST]

Pour finir la réalisation du transistor de sélection vertical, le surplus de Polysilicium est gravé afin de revenir au niveau de l'active. Cette gravure est critique, car si elle est insuffisante, des filaments de Polysilicium resteront le long des points mémoires, risquant de créer des courts-circuits. En revanche, si la gravure est trop importante il sera impossible de connecter la grille du transistor de sélection. Pour s'affranchir de l'affaissement du Polysilicium au niveau de la tranchée, une couche de BARC est déposée pour planariser la surface du Polysilicium. Puis une première gravure non sélective est utilisée pour graver le BARC aussi rapidement que le

Polysilicium afin d'arriver jusqu'en dessous de l'affaissement. Ensuite une gravure sélective est effectuée pour s'arrêter au niveau de la zone active (Figure 2.9).

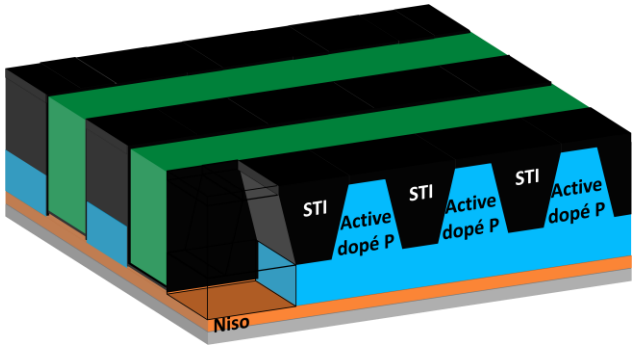


Figure 2.9 – Résultats après gravure du surplus du polysilicium

L'oxyde sacrificiel peut alors être retiré par une solution d'acide fluorhydrique (HF) et remplacé par un oxyde tunnel (Figure 2.10a). Cette étape est importante, car l'oxyde sacrificiel présente des impuretés suite aux différentes étapes qu'il a subies précédemment. Alors que l'oxyde tunnel doit être le plus propre possible, car c'est à travers ce dernier que les charges vont transiter entre le substrat et la grille flottante. Il aura aussi pour rôle de retenir les charges stockées lors de la phase de rétention. Le Polysilicium est ensuite déposé sur la totalité du wafer. Une étape de photolithographie (dépôt résine + masque de grille), présentée en Figure 2.10b, est effectuée pour définir la grille flottante (Figure 2.10c).

Figure 2.10 – a) Dépôt de l'oxyde tunnel b) Définition de la grille flottante c) Gravure de la grille flottante

Le diélectrique tri-couche ONO (Oxyde, Nitrure, Oxyde), est déposé au-dessus du polysilicium 1 (Figure 2.11a). Le polysilicium 2 est ensuite déposé sur tout le wafer (Figure 2.11b). À cet instant du procédé de fabrication, toutes les couches nécessaires à la réalisation des points mémoires sont présentes. Il reste uniquement à définir la longueur de grille souhaitée.

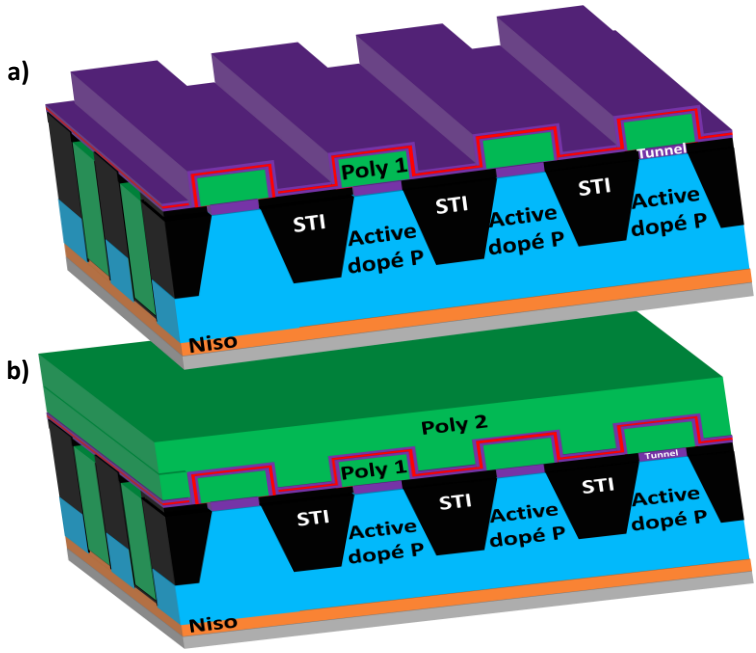


Figure 2.11 – a) Dépôt de la tri couche ONO b) Dépôt du Polysilicium 2 (grille de contrôle)

Pour cela une nouvelle étape de photolithographie (Figure 2.12a) est utilisée pour définir ces motifs. À l'aide d'une gravure en plusieurs étapes qui élimine successivement les couches Poly2/ONO/Poly1, il est possible de définir l'empilement mémoire (Figure 2.12b). Cette gravure particulièrement importante doit être maîtrisée, car elle définit la géométrie finale de la cellule mémoire. Les zones de source/drain sont alors implantées de part et d'autre de l'empilement de grilles des transistors par des implants de type N (Figure 2.12c).

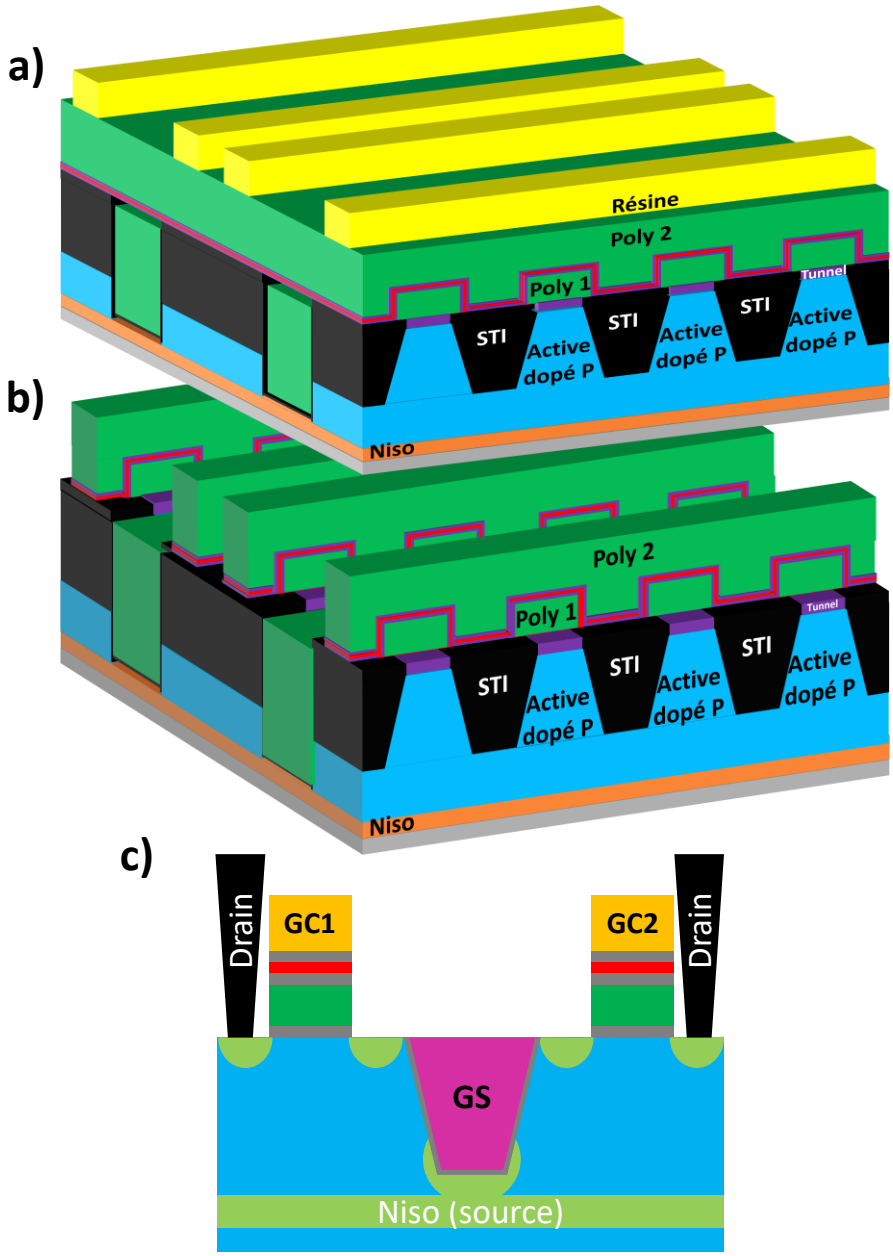


Figure 2.12 – a) Définition du point mémoire b) Gravure de l'empilement mémoire c) Coupe le long de l'active

L'étude du procédé de fabrication de l'eSTM étant terminée, il est important de noter que la difficulté de ce procédé est liée à la réalisation du transistor vertical qui est une première pour STMicroelectronics-Rousset. En plus des criticités de la partie mémoire s'ajoute la difficulté d'intégrer la partie logique[3]. La prochaine partie permettra de comprendre le comportement électrique de la cellule eSTM.

[3] Partie logique : c'est la partie qui permet le contrôle de la mémoire embarquée

II. Fonctionnement de la cellule eSTM

1. Présentation du banc de mesure

La caractérisation électrique de la cellule eSTM a été effectuée à l'aide d'une station sous pointe manuel (Prober). Ce Prober est piloté avec un programme sous Python qui permet de commander les instruments du banc d'essai (Figure 2.13). En particulier ce banc d'essai est équipé :

- D'un analyseur de paramètres de semi-conducteurs B1500 d'Agilent qui permet de générer les tensions à appliquer et de récupérer les courants via les quatre voies SMU[4].
- Deux matrices 16440A permettant de commuter entre les entrées/sorties SMU (servant à la lecture du point mémoire ou statique) et SPGU[5] (servant à générer les pulses haute tension de programmation et d'effacement).
- Quatre boîtiers RSU[6] qui permettent de réaliser les commutations des WGFMU[7] (permettant la mesure de la consommation durant la programmation) et les signaux des matrices (c'est à dire soit aux SMUs soit aux SPGUs).
- D'une station sous pointes 300mm (Prober)

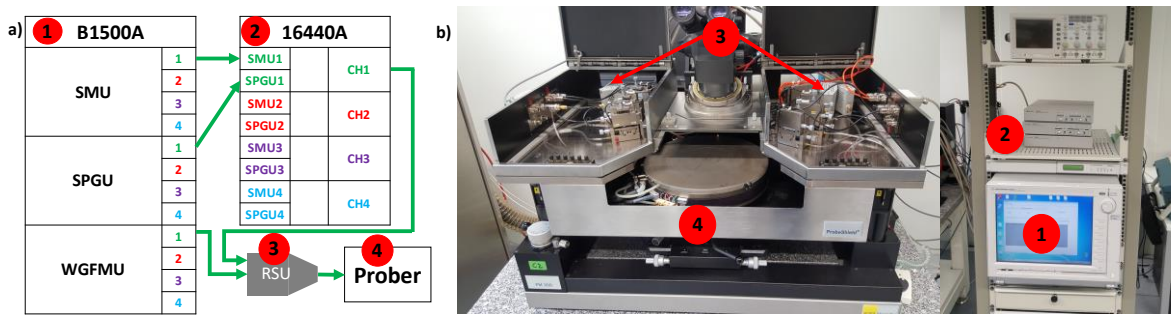


Figure 2.13 : Montage utilisé pour la caractérisation de la cellule eSTM

Ces mesures ont été effectuées au sein de l'équipe mémoire de l'IM2NP à Marseille. Dans la partie suivante nous allons présenter le principe de fonctionnement la cellule eSTM.

2. Le principe de fonctionnement de l'eSTM

La cellule eSTM est identique à une Flash NOR avec un transistor de sélection additionnel. La programmation se fait par le mécanisme SSI (chapitre 1 paragraphe III.2) comme pour une mémoire Split Gate. Ainsi la programmation de cette cellule est effectuée en appliquant une forte tension positive d'environ 10V sur la Grille de Contrôle (GC1) tandis qu'une tension de 1.7V est appliquée sur la Grille de transistor de Sélection (GS) afin d'ouvrir son canal. Une tension d'environ 4.2V est appliquée sur le drain afin de générer les porteurs chauds dans le canal du transistor à grille flottante. La source ainsi que la grille du transistor mémoire non sélectionné

[4] SMU : (Source Measure Unit) permet d'appliquer ou de mesurer des tensions / courants continus.

[5] SPGU : (Semiconductor Pulse Generator Unit) est utilisé pour réaliser des pulses de tension avec une résolution de 10ns avec des amplitudes de +/- 40V.

[6] RSU : Remote-sense & Switch Unit

[7] WGFMU : (Waveform Generator / Fast measurement Unit) permet de mesurer le courant de façon dynamique avec une résolution de 5ns.

(GC2) sont reliées à la masse afin de bloquer le point mémoire pour ne pas programmer la cellule non sélectionnée (Figure 2.14). On peut noter que la distance entre le transistor de sélection et les points mémoires a un impact sur les implants flottants dans cette zone. Une modification de la quantité et de la profondeur de ces implants peut avoir un impact sur le bon fonctionnement de la cellule mémoire, comme nous le verrons par la suite.

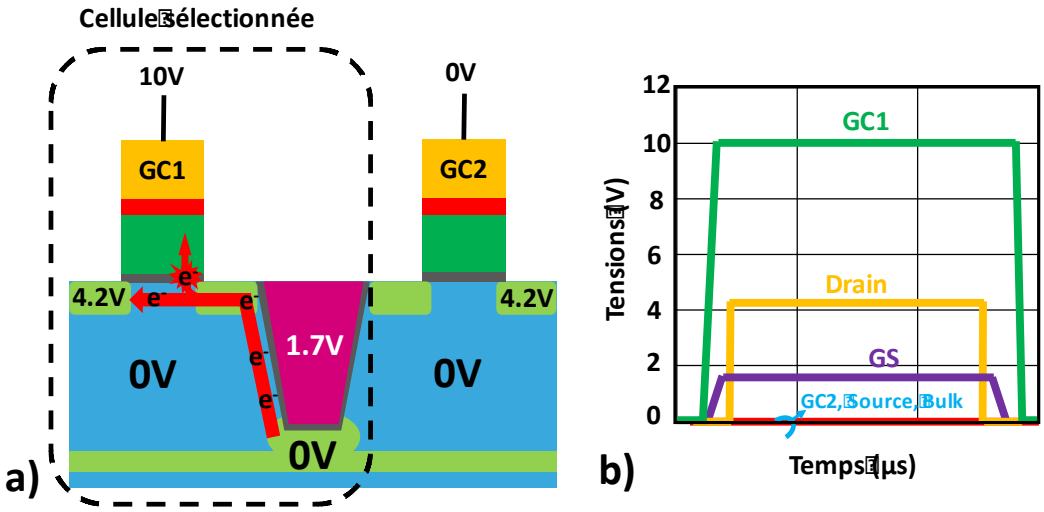


Figure 2.14 : a) Programmation par porteurs chauds et b) Chronogramme des signaux appliqués en programmation

L'effacement est réalisé par mécanisme FN, une tension de -10V est appliquée sur la GC1 ainsi qu'une tension de 8V sur le substrat afin de créer une forte différence de potentiel aux bornes de l'oxyde tunnel rendant possible l'effacement FN. Le drain est laissé flottant pour ne pas consommer de courant durant cette phase. La grille de transistor de sélection est polarisée à 8V afin de ne pas stresser l'oxyde de la tranchée qui voit une tension du substrat de 8V. Sur la grille GC2, une tension de 0V est appliquée pour diminuer la différence de potentiel aux bornes de l'oxyde tunnel afin d'éviter un effacement non désiré de la cellule non sélectionnée (Figure 2.15).

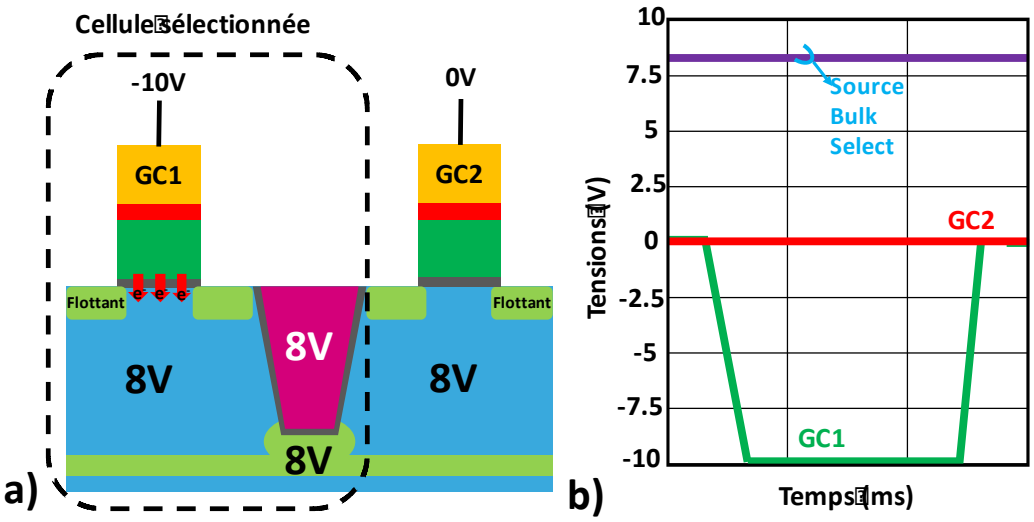


Figure 2.15 : a) Effacement par FN et b) Chronogramme des signaux appliqués en effacement

Afin de lire le point mémoire, le transistor de sélection est ouvert (V_{GS}=3V) laissant passer un courant de drain grâce à une tension de drain de 0.7V. La tension appliquée sur la grille du transistor mémoire sélectionné augmente jusqu'à ce qu'un courant de 3µA soit atteint. Le transistor mémoire non sélectionné est quant à lui bloqué en appliquant une tension de -3V sur sa grille GC2 (Figure 2.16).

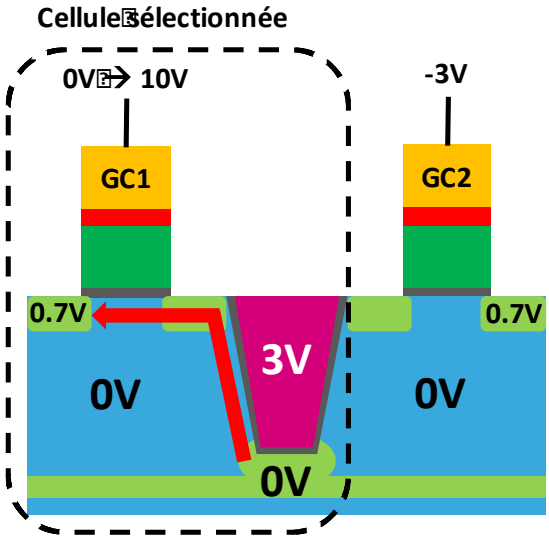


Figure 2.16 : Schéma de la cellule eSTM lors de la lecture

Dans le but de tester le bon fonctionnement d'une cellule mémoire, un test d'endurance est effectué à la fin du procédé de fabrication. Ce test consiste à observer l'évolution des tensions de seuils programmées (V_{TP}) et effacées (V_{TE}) après un grand nombre de cycles de programmation et d'effacement (Figure 2.17). En effet, ces deux niveaux vont avoir tendance à se rapprocher l'un de l'autre avec la dégradation de l'oxyde de tunnel. La cellule est considérée comme non fonctionnelle quand l'écart entre les deux tensions est insuffisant pour différencier l'état programmé de l'état effacé.

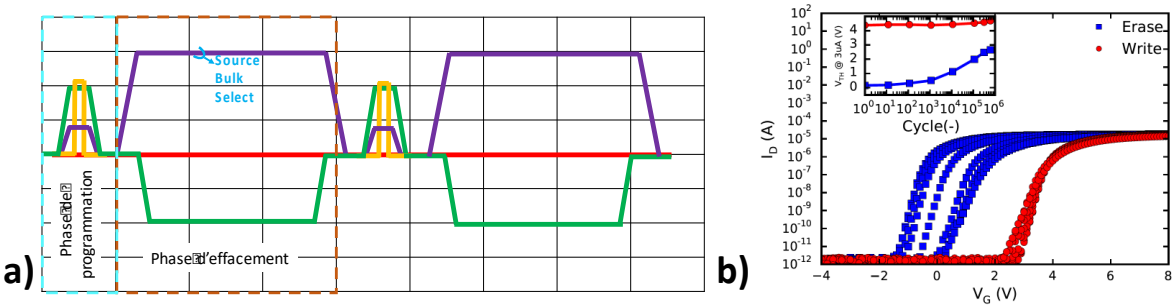


Figure 2.17 : a) Chronogramme de programmation et d'effacement utilisé durant la procédure d'endurance de la cellule. b) Caractéristique I-V de lecture d'une cellule eSTM au cours d'un test d'endurance. Insert : évolution des tensions de seuil programmé et effacé en fonction du nombre de cycles vu par la cellule.

III. Caractérisation électrique de la cellule eSTM

Dans cette partie nous allons étudier les performances électriques de la cellule eSTM. La première caractérisation va porter sur la spécificité de l'architecture proposée, à savoir le transistor vertical. Par la suite, la caractérisation du transistor mémoire sera présentée.

1. Caractérisation du transistor vertical

Dans cette partie l'étude du transistor vertical est présentée, plus particulièrement la caractéristique $I_D(V_{GS})$. Dans l'architecture eSTM, une seule grille est partagée par deux transistors mémoire : un pour le point mémoire du côté pair et l'autre du côté impair. Pour permettre la caractérisation d'un seul transistor de sélection une tension supérieure à 5V est appliquée sur la grille de contrôle de l'empilement mémoire le rendant passant. L'autre empilement est maintenu bloqué par l'application d'une tension de -3V. Toutefois, le positionnement du transistor mémoire dans un état programmé permet une meilleure sélectivité, limitant ainsi les tensions appliquées sur la grille de contrôle.

La Figure 2.18a illustre les différentes tensions appliquées sur la cellule eSTM pendant cette mesure. La caractéristique $I_D(V_{GS})$ du transistor vertical (Figure 2.18b) présente un V_T@1µA est de 1.45V. De plus, aucune différence n'est observée sur la mesure du courant entre le transistor de sélection côté pair ou impair. Ainsi les transistors verticaux présentent les caractéristiques nécessaires en tant que sélecteur mémoire. La prochaine étape consiste en la caractérisation du transistor mémoire.

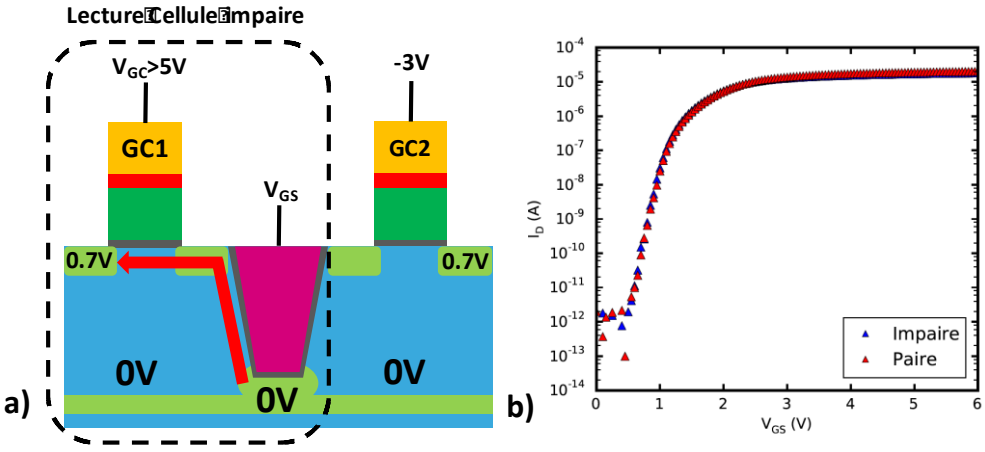


Figure 2.18 : a) Schéma des tensions appliquées durant la mesure – b) Caractéristiques $I_D(V_{GS})$ du transistor de sélection pour chaque transistor mémoire

2. Fenêtre de programmation de la cellule eSTM

Dans cette partie, le bon fonctionnement de la nouvelle cellule mémoire va être vérifié. Pour cela, des tensions de programmation et d'effacement proches de celles utilisées pour la cellule Flash vont être utilisées. Pour la programmation une tension de 4.2V sera appliquée sur le drain, 10V pour la grille de contrôle de l'empilement mémoire et 1.7V pour la grille du transistor vertical. Le reste est à la masse. Ces conditions sont décrites dans Figure 2.14. En ce qui concerne l'effacement, la tension de la grille de contrôle est fixée à -18V, le drain est considéré flottant et les autres tensions sont à la masse.

La Figure 2.19 représente les états programmé et effacé des cellules paires et impaires. Les fenêtres de programmation des cellules paires et impaires sont quasiment identiques avec une différence de seulement 100mV. Dans la suite de ce manuscrit, l'impact des tensions de programmation et d'effacement sur les performances de la cellule en termes de consommation et de fenêtre de programmation sera présenté.

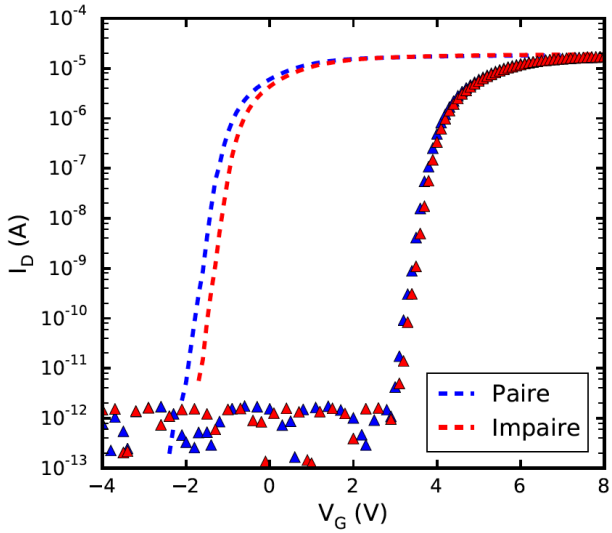


Figure 2.19 : Caractéristiques $I_D(V_{GC})$ du transistor mémoire pair et impair

3. Consommation de courant de la cellule eSTM

L'étude de la consommation de la cellule eSTM a été effectuée dans le cadre des travaux de thèse de J. BARTOLI [56]. Cette mesure de consommation repose sur l'utilisation de l'équipement Agilent B1500 et de son module WGFMU comme présenté dans la Figure 2.13. Le montage utilisé permet la mesure du courant de drain en dynamique avec une résolution de 10ns pendant l'application des signaux de programmation permettant ainsi l'extraction de l'énergie consommée par la cellule. Les tensions appliquées aux bornes de la cellule sont rappelées sur la Figure 2.20b. Elles sont appliquées sur la cellule impaire puis sur la cellule paire afin de comparer les courants de drain dynamique de ces deux cellules (Figure 2.20a).

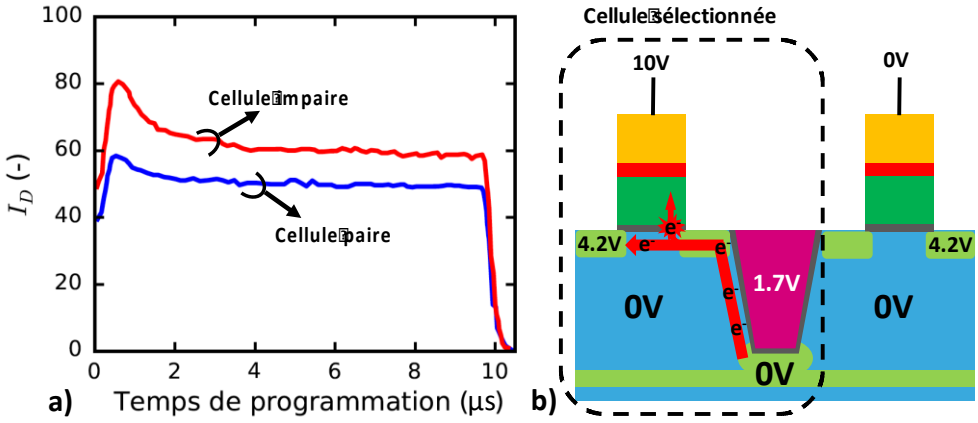


Figure 2.20 : a) Courant de consommation de la cellule paire et impaire – b) Polarisation de la cellule mémoire durant la phase de programmation

Contrairement aux fenêtres de programmation qui avaient un comportement symétrique, une différence non négligeable de la consommation entre les deux cellules paire et impaire est observée. Cette différence est probablement due à une dissymétrie de l'architecture eSTM comme l'expose la Figure 2.21. Cette dissymétrie est éventuellement due soit à la rugosité de la grille mémoire, soit à un désalignement du masque photolithographie de la grille mémoire et celui de la tranchée. Dans les deux cas, une différence sur l'implant flottant situé entre les transistors à grille flottante et le transistor de sélection modifiera le fonctionnement des deux cellules. La caractérisation de plusieurs lots a montré que ce n'est pas toujours la même cellule (paire ou impaire) qui fonctionne le mieux, ce qui illustre une variabilité lot à lot de la mesure d'alignement entre le réticule du transistor de sélection et celui de l'empilement mémoire.

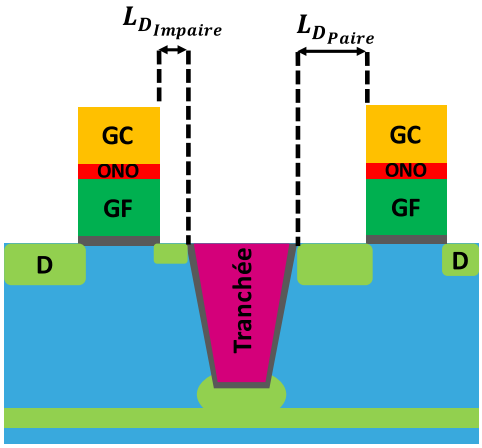


Figure 2.21 : Schéma d'une cellule eSTM désalignée

4. Energie totale consommée de la cellule eSTM

L'énergie totale consommée E_T durant la programmation de la cellule eSTM est l'image directe du courant de drain I_D. Il est possible de la calculer avec l'équation suivante :

$$E_T = \int_0^{t_p} I_D V_D \, dt$$

Avec t_p le temps de programmation

La Figure 2.22 montre les résultats de l'énergie totale consommée pour les cellules paire et impaire dans le cas de la Figure 2.20a. On remarque que l'énergie de consommation de la cellule paire est plus faible de 19% comparée à celle de la cellule impaire, en accord avec la mesure de courant faite auparavant.

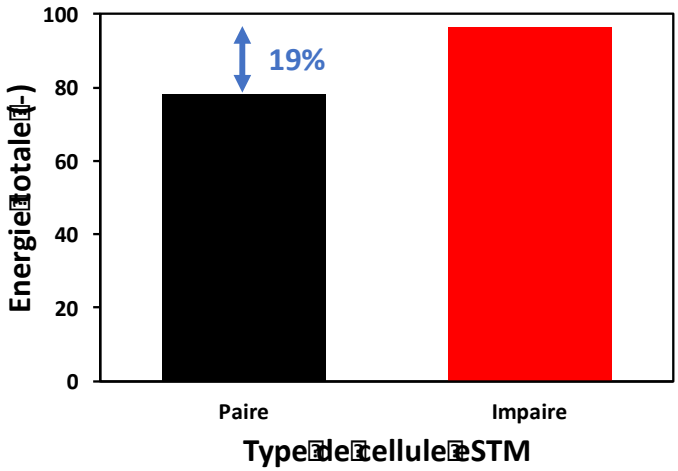


Figure 2.22 : Energie totale des cellules paire et impaire

5. Efficacité en programmation de la cellule eSTM

L'efficacité en programmation, quant à elle, est définie comme le rapport entre la fenêtre de programmation et l'énergie totale consommée.

$$E_f = \frac{V_{TP} - V_{TE}}{E_T}$$

Les fenêtres de programmation étant identiques pour les cellules paires et impaires, l'écart en efficacité sera uniquement dû à la différence de consommation. La Figure 2.23 représente l'efficacité des deux cellules, grâce à un courant de programmation plus faible la cellule paire est plus efficace de 21%. Ceci est dû à la dissymétrie de l'implant flottant des cellules mémoires.

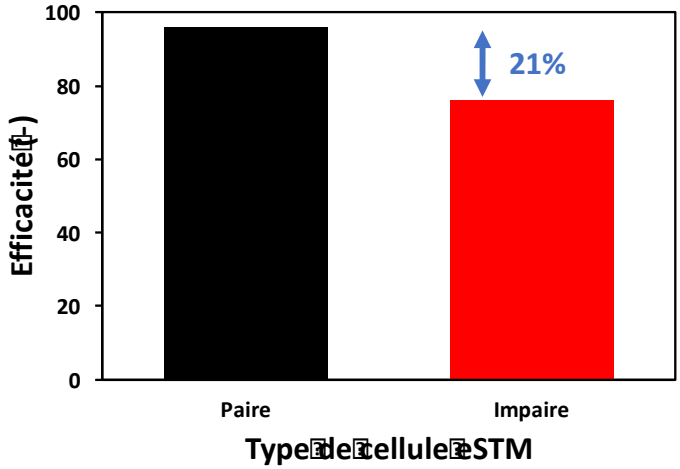


Figure 2.23 : Efficacité de la cellule paire et impaire

6. La fiabilité de la cellule eSTM

Usuellement, il existe deux critères pour évaluer la fiabilité d'une mémoire non volatile, à savoir l'endurance et la rétention d'information.

a) L'endurance de l'eSTM

L'endurance d'une cellule mémoire représente l'évolution de sa fenêtre de programmation après un certain nombre de cycles de programmation/effacement. Dans notre cas, 500k cycles sont appliqués. La Figure 2.24 met en avant les résultats obtenus sur des cellules paires et impaires, un comportement identique est observé. La tension de seuil V_{TP} reste identique même après 500k cycles. Cependant, la tension de seuil V_{TE} se dégrade en passant de 0V à plus de 2V (Figure 2.24a). Cette dégradation réduit la fenêtre de programmation de 45% (Figure 2.24b). Une étude visant à expliquer l'origine de la dégradation du V_{TE} sera présentée dans le paragraphe III.7d de ce chapitre.

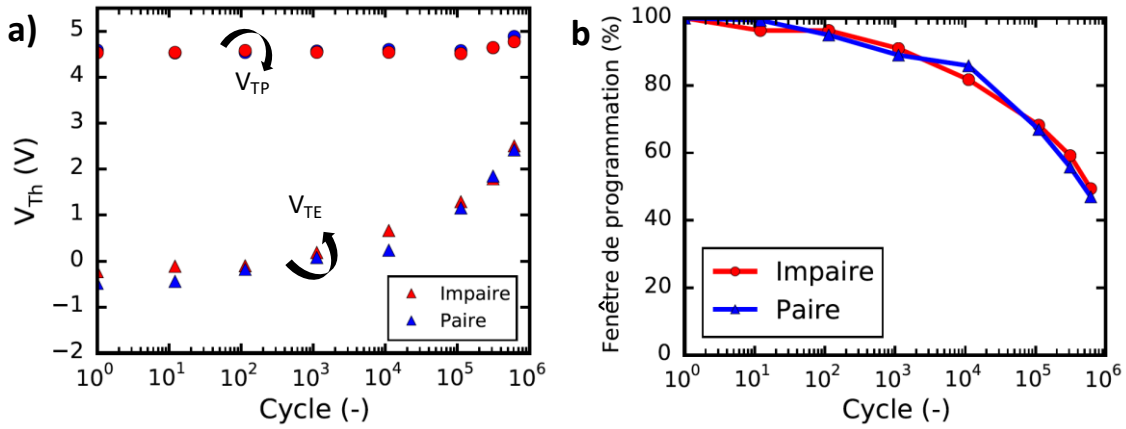


Figure 2.24 : a) Courbe d'endurance après 500k Cycles – b) Evolution de la fenêtre de programmation

b) Rétention de la cellule eSTM

Le but des tests en rétention est d'observer l'évolution de la perte de charges stockée dans la grille flottante au cours du temps. En se basant sur le cahier des charges des précédents produits, la mémoire eSTM doit conserver une fenêtre de programmation de 3V (soit une perte de 1.5V) au bout de 10 ans à 85°C. Pour accélérer la qualification des produits, les tests en rétention exploitent l'activation en température de la perte de charges qui suit une loi d'Arrhenius [57] dans les mémoires Flash. L'équation suivante basée sur la loi d'Arrhenius permet de calculer le facteur d'accélération (A_F) de cette perte de charges :

$$A_F = \exp\left(-\frac{E_a}{k}\left(\frac{1}{T_1}-\frac{1}{T_2}\right)\right)$$

Où :

A_F : Facteur d'accélération
E_a : Energie d'activation (0,6eV)
T_1 : Température de départ en kelvins
T_2 : Température de tests en kelvins
k : Constante de Boltzmann (8,62 x 10^{-5} eV.K^{-1})

Ainsi une rétention de 10 ans à 85°C est équivalente à 6 mois (4400h) à 150°C ou encore 8 jours (192 heures) à 250°C. La Figure 2.25 présente les mesures en rétentions à 150°C et à 250°C réalisées par l'équipe de caractérisation électrique de STMicroelectronics-Rousset. Elle montre qu'au bout de 192 heures à 250°C, la perte de V_{TP} ne dépasse pas 0.2V, ce qui reste bien inférieur au maximum toléré de 1.5V. Cette expérience illustre que la rétention de la cellule eSTM est au-dessus des limites de fonctionnement à garantir.

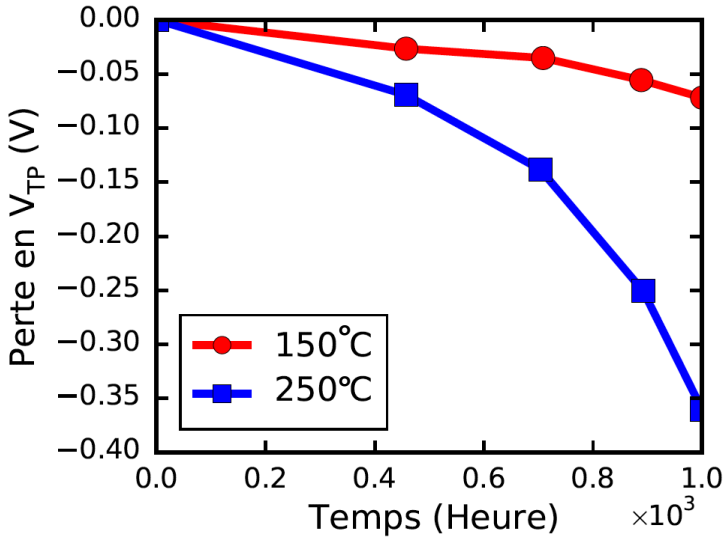


Figure 2.25 : Courbes de rétention mesurées sur des cellules eSTM pour deux températures (150°C et 250°C)

7. Comparaison entre la cellule eSTM et une cellule Flash

a) Fenêtre de programmation

Les mesures présentées dans cette partie ont été effectuées à STMicroelectronics-Rousset, la Figure 2.26 représente les cinétiques de programmation des deux cellules mémoires : eSTM et Flash. Pour une durée de 10µs la cellule Flash atteint une tension de seuil V_{TP} de 7V contrairement à l'eSTM qui sature dès 4V. Dans un souci de positionner les deux cellules mémoires dans un état similaire afin de comparer leurs performances, leurs fenêtres de programmation ont été fixées à 4V. Ainsi la mémoire Flash atteint cette fenêtre de programmation en 0,8 µs tandis que l'eSTM l'atteint au bout de 4,2 µs. Pour la suite nous étudierons les performances de chacune de ces mémoires pour leurs temps de programmation respectifs.

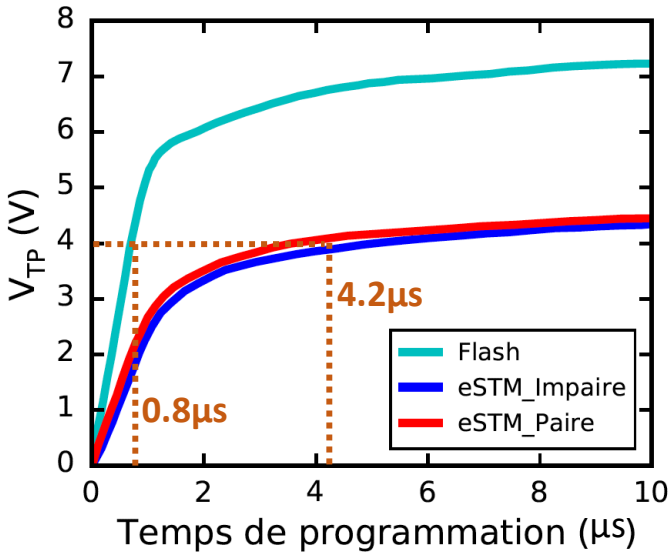


Figure 2.26 : Cinétique de programmation d'une cellule Flash et des cellules eSTM paire et impaire.

b) Consommation

Les courants de consommation de chaque cellule mémoire sont représentés dans la Figure 2.27 pour une durée de programmation totale de 10 µs. Cependant, les zones colorées sous ces courbes représentent les courants de consommation de chaque cellule pour les durées de programmations nécessaires afin d'atteindre une fenêtre de programmation de 4,5V.

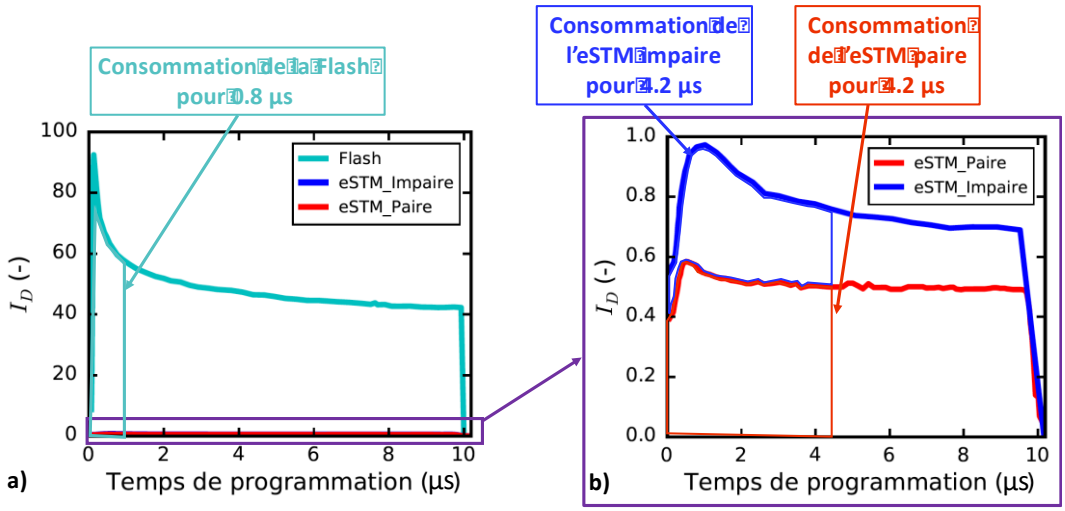


Figure 2.27 a) Courant de consommation d'une cellule Flash – b) Courant de consommation des cellules eSTM paire et impaire.

Si nous comparons les courants de programmation des cellules Flash ou eSTM, nous constatons que la cellule eSTM nécessite beaucoup moins de courant que la cellule Flash pour être programmée (avec une réduction d'environ 90% du courant de programmation).

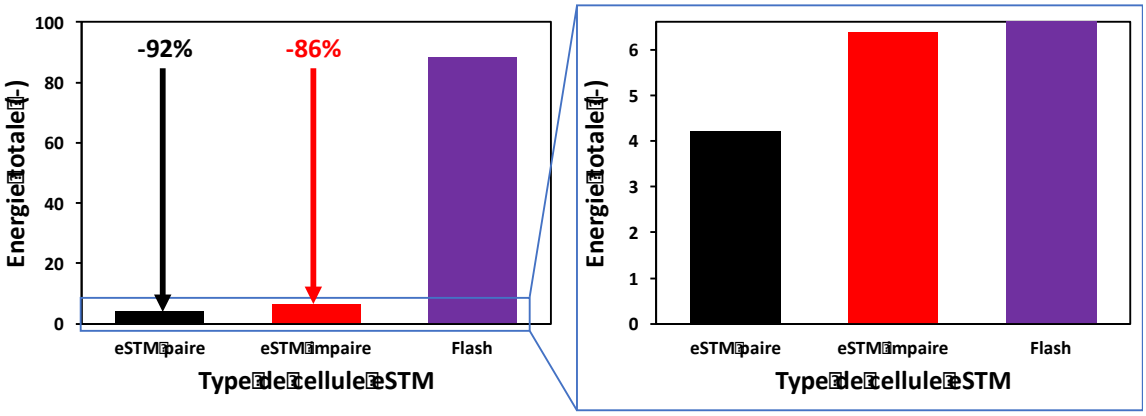


Figure 2.28 : Energie totale de consommation d'une cellule Flash et des cellules eSTM paire et impaire

La Figure 2.28 illustre le gain d'environ 90% sur l'énergie totale consommée entre la Flash et les cellules eSTM. De plus la Figure 2.29 souligne que l'efficacité de programmation des cellules eSTM est en moyenne 20 fois plus importante qu'une cellule Flash standard. Ainsi la cellule eSTM présente véritablement sa supériorité face aux cellules Flash standard pour les applications ultra-basse consommation.

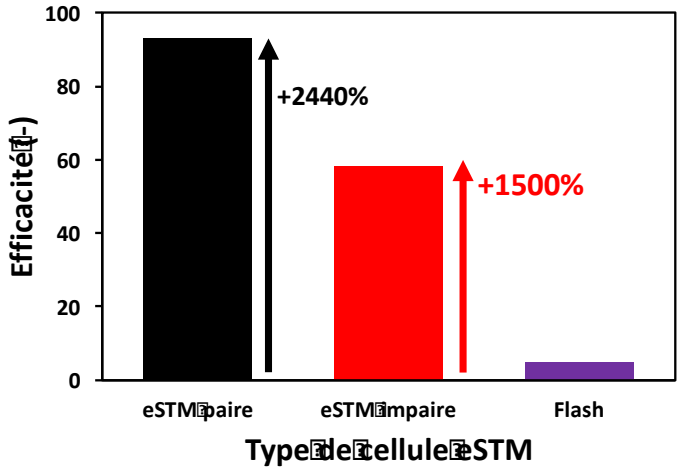


Figure 2.29 : Efficacité de programmation d'une cellule Flash et des cellules eSTM paire et impaire

c) Endurance

Les mesures d'endurance réalisées par STMicroelectronics sur la cellule Flash utilise un temps de programmation de 10 µs, et non pas 0,8 µs, ainsi la tension de seuil de l'état programmé sera de 5V. La Figure 2.30a illustre l'évolution des tensions de seuil V_{TP} et V_{TE} et celle de la fenêtre de programmation (Figure 2.30b) en fonction des cycles successifs de programmation / effacement (500k cycles). Au début du test, jusqu'à 1000 cycles, les deux cellules se comportent de façon similaire en gardant la fenêtre de programmation constante. Au bout de 500k cycles la fenêtre de programmation de l'eSTM et celle de la Flash sont réduites de 46%, soit une perte d'environ 2V pour l'eSTM et de 2.5V pour la cellule Flash.

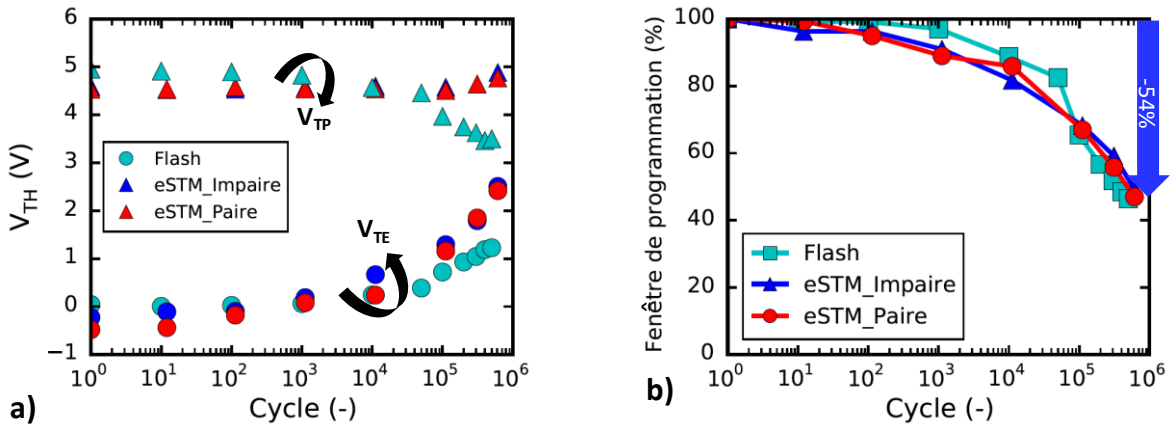


Figure 2.30 : a) Evolution des tensions V_{TP} et V_{TE} au bout de 500k cycles - b) Evolution de la fenêtre de programmation pour les cellules Flash, eSTM impaire et paire

Au vu des résultats en endurance présentés dans la Figure 2.30, nous remarquons que le comportement de la fenêtre de programmation de la cellule eSTM est différent de celui de la Flash. Dans le cas de la mémoire Flash, la fermeture de la fenêtre de programmation est liée à la dégradation des tensions de seuil V_{TE} et V_{TP}. Alors que pour l'eSTM, la dégradation provient essentiellement d'une remontée de la tension de seuil de l'état effacé.

Il est intéressant de constater que le comportement observé de notre cellule Flash correspond bien à celui constaté sur d'autres tests similaires présents dans la littérature comme présenté dans la Figure 2.31a [58]. En ce qui concerne la cellule eSTM, son comportement est très similaire à celui des mémoires Split Gate [59], cf. La Figure 2.31b.

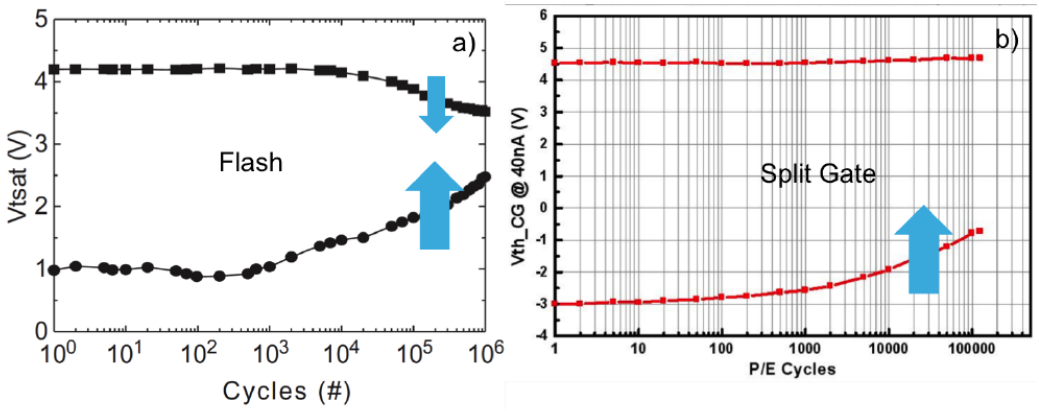


Figure 2.31 : Exemples de courbes d'endurance d'une cellule Flash [58] **et d'une cellule Split Gate** [59]

Dans la partie suivante, nous allons proposer une explication de la dégradation de la fenêtre de programmation de la cellule eSTM.

d) Localisation des défauts

La dégradation de la fenêtre de programmation eSTM est due principalement à une augmentation de la tension de seuil de l'état effacé. Nous supposons que les défauts peuvent se situer le long du canal du transistor de sélection et / ou le long de celui du transistor mémoire (Figure 2.32).

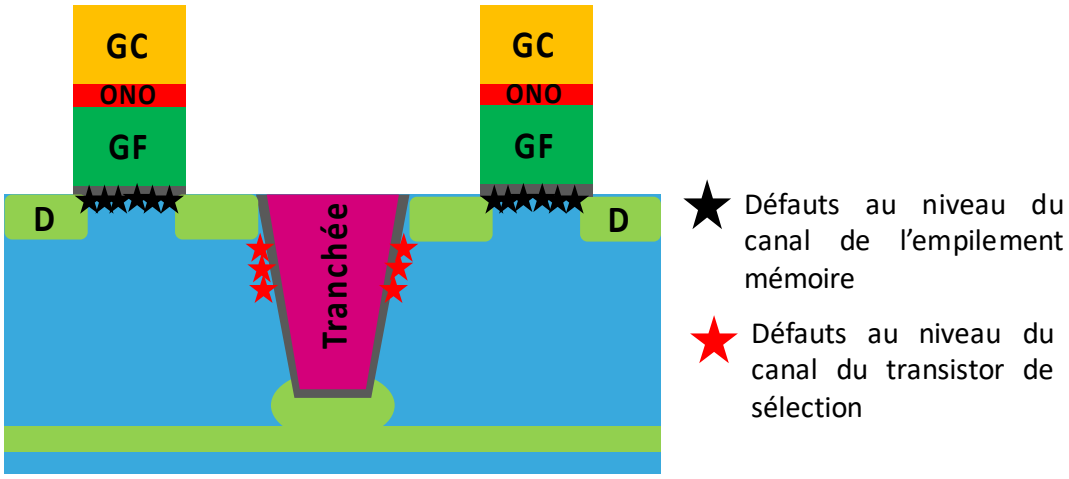


Figure 2.32 : Schéma explicatif de la dégradation de la cellule eSTM après un test d'endurance

Pour vérifier l'apparition des défauts au niveau du canal de la tranchée, une mesure $I_D(V_{GS})$ a été exécutée avant et après un test d'endurance. L'objectif est d'observer une dégradation de la pente sous le seuil. La Figure 2.33 montre que la pente sous le seuil de la caractéristique $I_D(V_{GS})$ reste identique. Cette mesure nous permet d'affirmer qu'il n'y a de création de défauts au niveau du canal du transistor de sélection après un test d'endurance de 500k cycles.

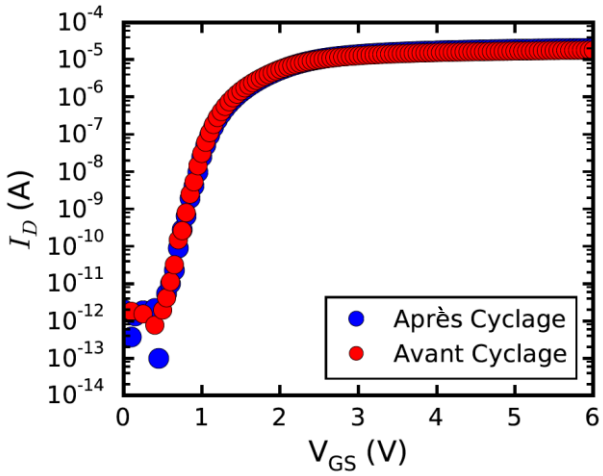


Figure 2.33 : Caractéristique $I_D(V_{GS})$ du transistor de sélection avant et après un test d'endurance de 500k cycles.

Comme observé dans la Figure 2.30a, la tension de seuil V_{TP} de l'eSTM augmente légèrement alors que celle de la Flash diminue. Ceci peut être expliqué par la méthode de programmation utilisée pour les mémoires Flash (i.d. par porteurs chauds). Cette augmentation dégrade localement leur oxyde tunnel à proximité du drain [60]. D'une part, ce mécanisme de programmation engendre des états d'interfaces qui vont diminuer la probabilité d'injection des électrons dans la grille flottante [61], [62]. D'autre part, des électrons piégés dans l'oxyde tunnel vont modifier le champ électrique au point d'injection ce qui va diminuer l'efficacité de programmation. Dans le cas de la cellule eSTM, l'injection des porteurs chauds se fait plus proche de l'implant flottant que du drain. L'analyse des courbes $I_D(V_G)$ de la cellule eSTM (Figure 2.34) fait ressortir une dégradation de la pente sous le seuil tout au long du test d'endurance. Ce phénomène indique l'apparition de défauts à l'interphase oxyde tunnel / canal [63]. De plus, le

décalage de la caractéristique $I_D(V_G)$ pour l'état effacé indique la création de pièges profonds chargés négativement limitant l'efficacité de l'effacement [63].

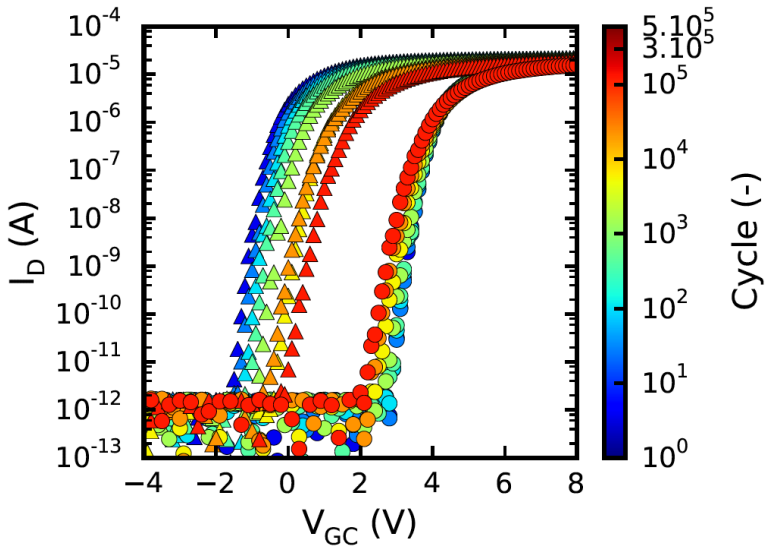


Figure 2.34 : Evolution des caractéristiques $I_D(V_{GC})$ au bout de 500k cycles

e) Evolution de la consommation après cyclage

La dégradation des performances de la cellule eSTM et celle de la Flash après quelques cycles seront présentées dans cette partie. L'évolution du courant de programmation en fin de vie, à savoir après 500k cycles, est illustrée dans la Figure 2.35.

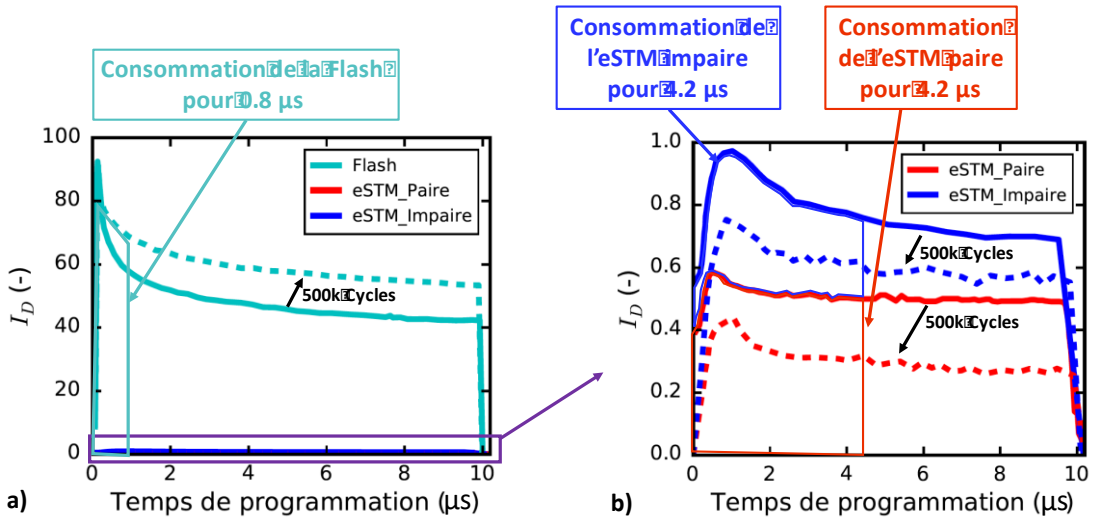


Figure 2.35 : a) Evolution du courant de consommation de la Flash après cyclage – b) Evolution du courant de consommation des cellules paire et impaire après cyclage.

Contrairement au courant de programmation de la mémoire Flash qui augmente après le test d'endurance (Figure 2.35a), celui de la cellule eSTM diminue après dégradation (Figure 2.35b). Le courant de programmation de la Flash augmente de 4% après 500k cycles (augmentation déjà montrée dans la thèse de G. Just [64]) tandis que celle de l'eSTM diminue de 30% en moyenne. Dans la littérature, l'augmentation ou la diminution de la consommation du

courant dynamique d'une cellule Flash après un test en endurance d'un million de cycle a été expliquée [65]. La Figure 2.36 provenant des travaux de V. Della Marca *et. al,* [65] met en avant l'impact des pièges générés lors de la programmation sur la consommation d'une mémoire Flash. Dans un premier temps, la mesure du courant dynamique est faite du côté stressé (Figure 2.36b). De cette façon, les pièges/charges négatives sont présents dans la zone d'injection des porteurs chauds réduisant ainsi leur injection et par conséquent le potentiel de la grille flottante diminue peu, laissant un courant fort traverser la cellule durant toute la durée de programmation. Dans un deuxième temps, la consommation dynamique est mesurée du côté non stressé (Figure 2.36c). Dans ce cas, la tension V_{TP} après cyclage équivaut à celle de la première programmation de la cellule, mais les charges piégées localement du côté stressé augmentent la tension V_{TP}. Par conséquent, le courant de drain diminue ce qui engendre une diminution de la consommation dynamique. Dans le cas de l'eSTM, l'énergie utilisée pour programmer la cellule est très importante maintenant constante la quantité de charges injectées dans la grille flottante durant le test d'endurance. Néanmoins les charges négatives piégées dans l'oxyde tunnel augmenteront au fil des cycles la tension de seuil durant la programmation limitant ainsi le courant de drain et par conséquent la consommation de programmation.

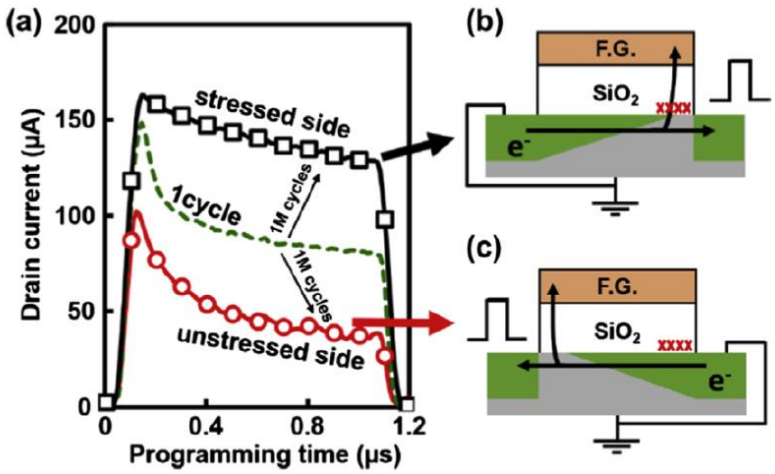


Figure 2.36 a) Courant de consommation d'une cellule Flash avant et après 1 million de cycles sur le côté stressé et non stressé. – b,c) Comportement de la cellule pendant la programmation.

La comparaison des caractéristiques de la cellule mémoire eSTM avec celle d'une mémoire Flash a montré essentiellement que la cellule eSTM est plus appropriée pour les applications ultra-basse consommation. Dans la prochaine partie l'évolution des caractéristiques de l'eSTM en fonction des différentes conditions de fonctionnement sera présentée.

8. L'évolution des caractéristiques de la cellule eSTM

Pour cette étude, nous conserverons les conditions précédentes de programmation et d'effacement de l'eSTM comme référence (Figure 2.14) et nous ne ferons varier qu'un seul des paramètres à la fois.

a) L'effet de V_{GC} pendant la programmation

Pour réaliser ces mesures, la tension de grille des points mémoire V_{GC} a été variée de 8V à 10V afin d'étudier l'évolution de la fenêtre de programmation ainsi que la consommation des deux cellules paire et impaire.

La Figure 2.37 illustre l'évolution de la fenêtre de programmation en fonction de la tension de grille du transistor mémoire pour les cellules paire et impaire. Cette figure montre aussi qu'il n'y a aucune différence au niveau de la fenêtre de programmation pour les deux cellules. En revanche, l'augmentation de la tension de grille de contrôle améliore la fenêtre de programmation. La tension V_{GC} a le même rôle pour l'eSTM que pour une cellule à grille flottante standard. Plus la tension de grille est élevée plus le champ vertical est important. Avec ces conditions les électrons chauds ont plus de chance d'être injectés dans la grille flottante augmentant ainsi la tension de programmation.

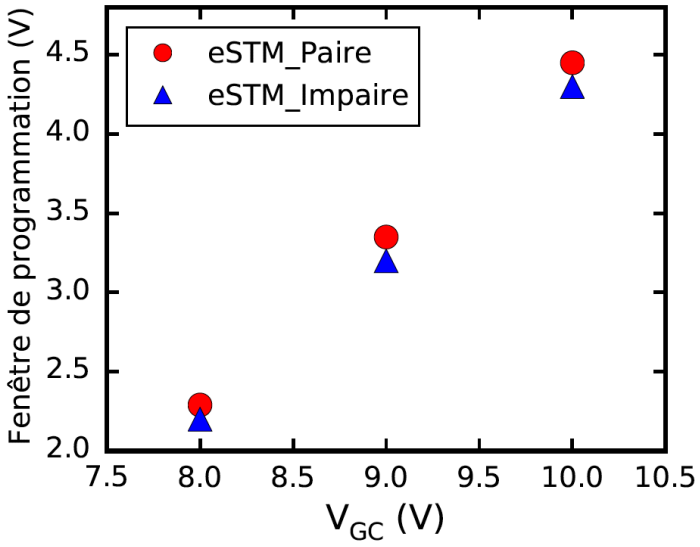


Figure 2.37 : Evolution de la fenêtre de programmation en fonction de la tension de la grille mémoire.

b) L'effet du V_D pendant la programmation

Pour l'étude de l'impact de la tension de drain sur les caractéristiques de la cellule eSTM pendant la programmation, cette tension a été variée de 3,6V à 4,6V par pas de 0,2V tout en gardant les autres tensions de références.

La Figure 2.38 montre l'évolution de la fenêtre de programmation en fonction de la tension de drain pour les deux cellules mémoires. Quel que soit le côté programmé, l'augmentation de la tension de drain génère une amélioration de la fenêtre de programmation. Pour garantir une fenêtre de programmation élevée, la tension de drain optimale serait 4,6V. Cependant, cette tension est proche de la tension de claquage de la jonction drain/substrat. Pour éviter une destruction des dispositifs pendant les tests d'endurance, une tension de drain de 4,2V a été choisie pour préserver la cellule eSTM.

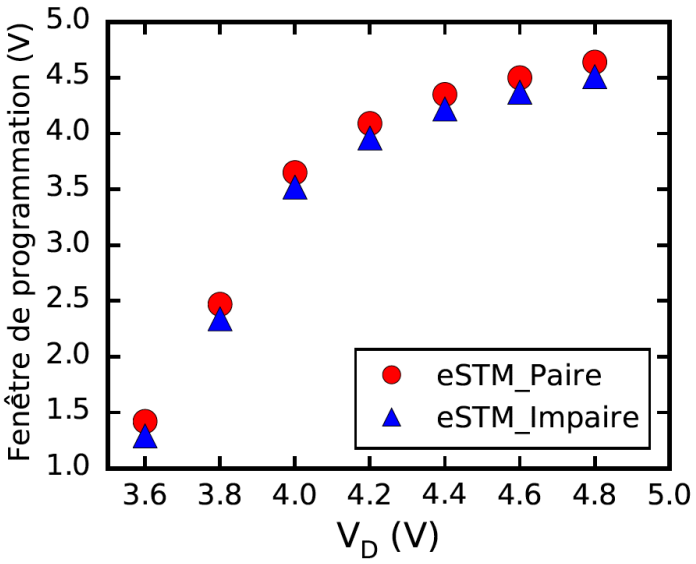


Figure 2.38 : Evolution de la fenêtre de programmation en fonction de la tension de drain.

c) L'impact de la tension de grille V_{GS}

Pour cette étude, la tension de la grille du transistor de sélection V_{GS} varie de 1,3V à 2,3V par pas de 0,2V. La Figure 2.39 représente la fenêtre de programmation en fonction de la tension de grille du transistor de sélection. Nous constatons que la fenêtre de programmation optimale est obtenue avec une tension V_{GS} aux alentours de 1,7V.

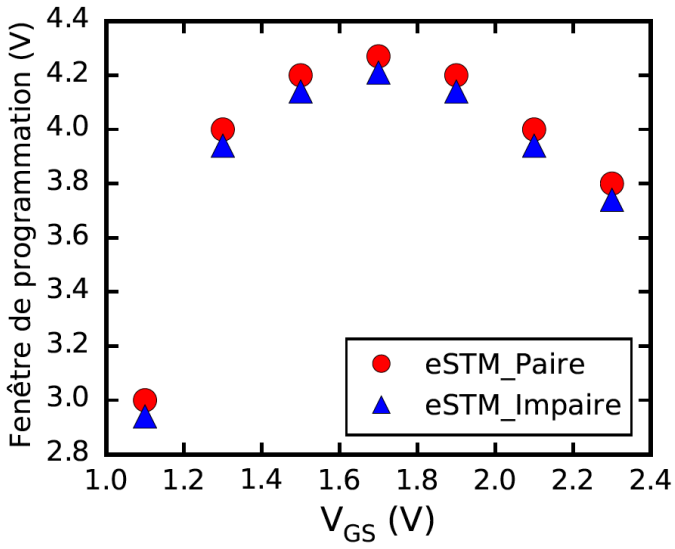


Figure 2.39 : Evolution de la fenêtre de programmation en fonction de la tension de grille du transistor de sélection

d) L'impact sur la consommation

La mesure du courant de drain durant la programmation pour les différentes conditions a été effectuée dans le cadre des travaux de J. Bartoli. L'impact de la variation de la tension de grille de contrôle sur ce courant a été représenté dans la Figure 2.24. Contrairement à la fenêtre de programmation (Figure 2.40), nous ne constatons aucune différence de courant de consommation pour ces trois tensions différentes. Cette indifférence à la tension V_{GC} peut être

expliquée par la présence du transistor de sélection qui impose le courant à travers la cellule mémoire inhibant ainsi l'effet de l'augmentation de la tension de grille sur le courant qui traverse la mémoire.

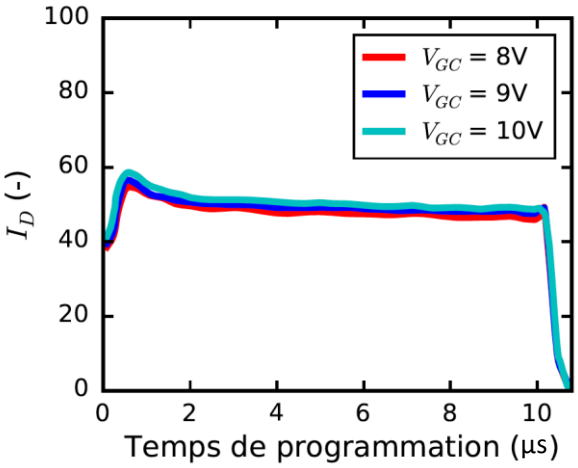


Figure 2.40 : Courant de programmation pour différentes tensions de grille de contrôle des points mémoires.

L'impact de la tension de drain sur les courants de consommation est représenté dans la Figure 2.41. Le courant de programmation reste le même pour les différentes conditions mis à part une légère différence au niveau du pic initial. Cette différence reste négligeable sur la durée totale de l'impulsion de 10µs.

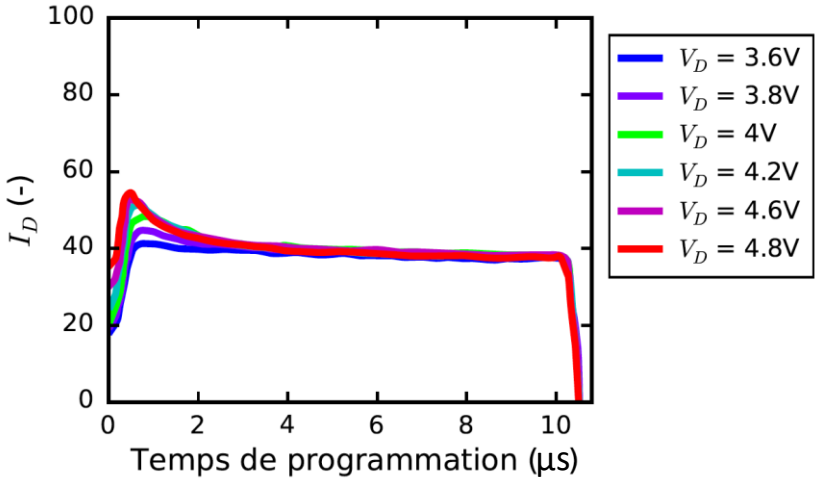


Figure 2.41 : Courant de programmation pour différentes tensions de drain

La Figure 2.42 représente les courants de programmation pour les différentes tensions du transistor de sélection. Une nette augmentation de ce courant est observée avec la croissance de de la tension V_{GS}.

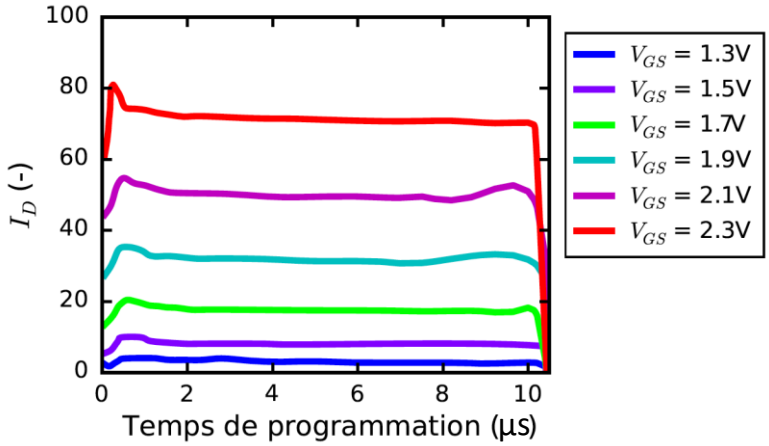


Figure 2.42 : Courant de programmation pour différentes tensions de grille du transistor de sélection

L'analyse des mesures effectuées dans cette partie montre que la tension de grille du transistor de sélection est celle qui a le plus fort impact sur le courant de consommation. Ces résultats confirment que la fonction principale du transistor de sélection est de moduler le courant qui traverse la cellule mémoire. Cette topologie, empilement mémoire associé à un transistor de sélection, comme les mémoires Split Gate ou la cellule mémoire eSTM, est la plus appropriée pour adresser les applications basse consommation.

La suite de ce manuscrit présente l'étude de l'impact du procédé de fabrication sur la fiabilité de la cellule eSTM et plus précisément sur l'endurance.

IV. Conclusion

Dans la première partie de chapitre, nous avons présenté la description du procédé de fabrication de l'architecture eSTM. Ensuite une étude complète du fonctionnement de la cellule eSTM a été présentée. Cette étude a mis en avant ses performances et sa faible consommation. Les caractéristiques de la cellule eSTM ont été comparées à celle d'une cellule Flash standard. Cette analyse a montré que la Flash se programme plus vite que l'eSTM. En revanche, la consommation de la cellule eSTM est drastiquement réduite par rapport à la cellule Flash (90%). Les résultats de cette partie prouvent que l'architecture eSTM est une bonne candidate pour concurrencer les mémoires de type Split Gate grâce à la simplicité de son procédé de fabrication.

Le prochain chapitre consiste à étudier la variabilité des étapes du procédé de fabrication et leur impact sur le bon fonctionnement électrique de la cellule eSTM.

Chapitre 3 : ETUDE DE VARIABILITE DU PROCEDE DE FABRICATION DE L'eSTM

Sommaire

Ce chapitre est dédié à l'étude de la variabilité des étapes du procédé de fabrication de la cellule eSTM. L'objectif principal est d'identifier les caractéristiques physiques qui sont à l'origine de la variabilité des paramètres électriques. Une importante campagne de mesures en ligne a été réalisée afin d'étudier la variabilité des paramètres physiques due au procédé de fabrication. Avant de décrire cette campagne de mesures et ses résultats, nous allons présenter dans un premier temps la méthodologie statistique utilisée et en justifier le choix. La seconde partie est consacrée à une étude approfondie des criticités du procédé de fabrication de l'eSTM. Nous verrons aussi les solutions apportées pour améliorer la variabilité des paramètres électriques.

I. Analyse de variabilité : Méthodologie

Avant d'introduire la méthodologie utilisée pour l'analyse de variabilité, nous devons définir le terme qui est au centre de ces travaux de thèse : la variabilité. La production dans l'industrie du semi-conducteur est réalisée par lot, chaque lot contient 25 plaques en silicium monocristallin. Dans ce cadre, la variabilité d'un paramètre, qu'il soit physique ou électrique, est spatiale et temporelle [66]. La variation spatiale est la somme d'une variation intra-plaque due, par exemple, à un dépôt par centrifugation non uniforme ou à un procédé de gravure dont la vitesse dépend du rayon, et d'une variation intra-champs, liée par exemple au procédé de photolithographie. La variabilité temporelle, lot à lot, ou plaque à plaque, est due principalement aux dérives des procédés de fabrication dans le temps.

1. Méthodologie

À partir de l'état de l'art des travaux sur ce même sujet, une méthodologie d'analyse générique [67] est appliquée en trois points :

- Une corrélation entre les paramètres électriques (tension de seuil, courant de drain …) et un ensemble de caractéristiques physiques mesurées en ligne (longueur de grille, épaisseur d'oxyde …) est effectuée à l'aide d'une régression multiple pas à pas ou une régression des moindres carrées (*Partial Least Squares* : PLS).
- Une identification des briques de procédé de fabrication associées aux sources de variabilité les plus importantes.
- Enfin, une analyse de variance pour dissocier les composantes spatiales des composantes temporelles, des briques du procédé identifiées au préalable.

L'application de cette méthodologie nécessite un grand nombre d'échantillons identiques. Dans un cas de production de masse cette méthodologie est amplement suffisante pour effectuer une analyse de variabilité. Cependant, le développement d'une nouvelle architecture mémoire nécessite beaucoup d'essais afin de trouver le point de fonctionnement de cette dernière. C'est pour cette raison qu'une méthodologie standard d'analyse de variabilité ne peut pas être appliquée sur la cellule eSTM par manque d'échantillons identiques.

Afin d'étudier la variabilité durant le développement de l'eSTM, une étude approfondie des étapes du procédé de fabrication qui sont critiques au bon fonctionnement de la cellule eSTM a été proposée. Pour orienter cette campagne de mesures des paramètres physiques, nous nous sommes concentrés sur l'étude théorique de cette cellule mémoire ainsi que sur les premiers résultats électriques.

2. Mesures des paramètres physiques

L'industrie des semi-conducteurs utilise deux méthodes de caractérisation des dimensions critiques, à savoir la scattérométrie et la microscopie électronique à balayage (SEMCD). Dans un environnement de production, il est important d'adopter l'outil de métrologie le plus adapté pour assurer une bonne précision et une bonne fiabilité.

La scattérométrie est une métrologie optique indirecte, fondée sur l'analyse de la lumière diffractée sur une structure de test (Figure 3.1). D'un point de vue matériel un scattéromètre est généralement une association de trois éléments : un éllipsomètre, un simulateur électromagnétique de la réponse scattérométrique et des modules de calculs très avancés [68]. La structure de test mesurée forme un réseau périodique de motifs (lignes, tranchées, etc.) permettant d'extraire le profile moyen de la dimension critique par diffraction. Cette technique, insensible aux variations locales, permet l'utilisation d'un spot lumineux (35µm) largement supérieur aux dimensions ultimes des technologies avancées. De plus, du point de vue simulation la construction de la réponse optique se trouve plus aisée dans le cas d'une structure périodique. Ces structures sont généralement réalisées dans les lignes de découpe rendant impossible la mesure scattérométrie au sein même du circuit. Ainsi cette technique n'a pas pu être utilisée durant la thèse, car les mesures devaient être effectuées directement sur le plan mémoire.

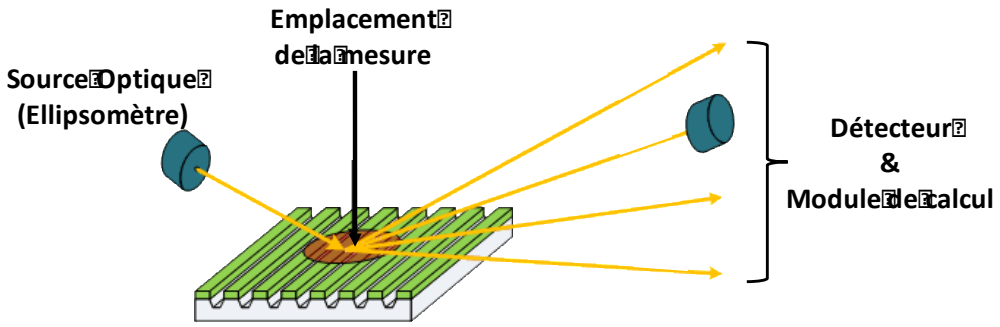


Figure 3.1 : Principe de fonctionnement de la scattérometrie. (Source : N&K Technology)

D'un autre côté, le microscope électronique à balayage est un outil classique utilisé dans l'industrie du semi-conducteur. Cet équipement permet de déterminer les dimensions critiques (CD) en balayant la surface avec un faisceau d'électrons. Les images produites sont formées par les électrons secondaires émis par l'échantillon au cours de son bombardement (Figure 3.2). L'équipement SEMCD est relativement rapide et complètement automatisé. N'exigeant aucune contrainte particulière en ce qui concerne la structure de mesure, il est très facile de mesurer directement dans le plan mémoire.

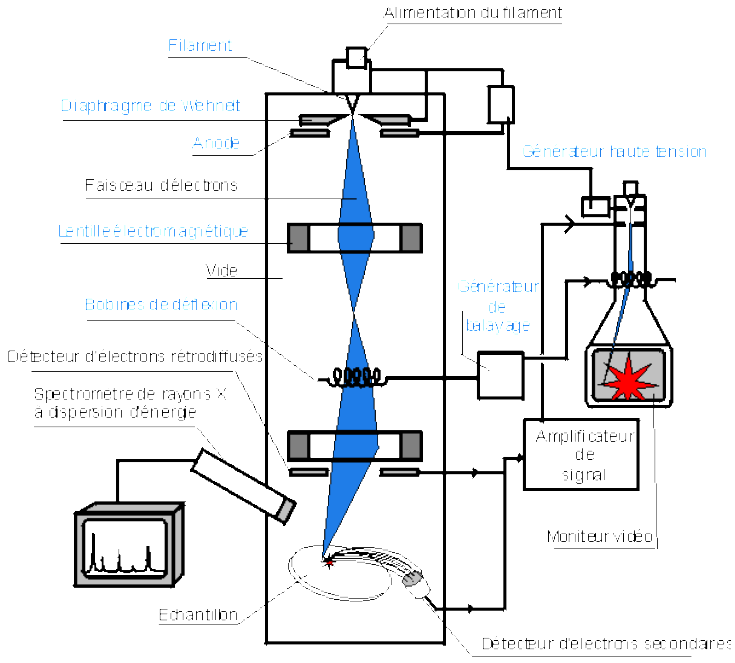


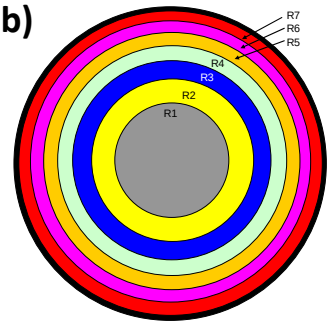
Figure 3.2 : Principe de fonctionnement du microscope électronique à balayage SEM [55].

Les mesures effectuées par le SEMCD sur une plaque de silicium respectent "la théorie des rings" (Figure 3.3). Cette approche a pour objectif de diminuer le temps de mesure en minimisant le nombre de points de mesure par opération. Dans le cas de STMicroelectronics-Rousset la mesure est effectuée sur neuf sites. Il faut noter que d'une manière générale le nombre de plaques mesurées par lot après une opération précise est limité à deux. Cet échantillonnage est utilisé pour limiter le coût final de la production.

Pour réaliser une étude de variabilité, nous avons opté pour un nombre de mesures plus important comparé au nombre de mesures standards. Ainsi, le nombre de sites mesurés sur chacune des plaques est relativement important (27 sites de mesure), et toutes les plaques des lots d'essais sont mesurées.

a)

Ring	Distance au bord de plaque
Ring 1	du centre à 63.3 mm
Ring 2	de 63.3 mm à 48.1 mm
Ring 3	de 48.1 mm à 36.5 mm
Ring 4	de 36.5 mm à 26.7 mm
Ring 5	de 26.7 mm à 18 mm
Ring 6	de 18 mm à 10.2 mm
Ring 7	de 10.2 mm à 3 mm



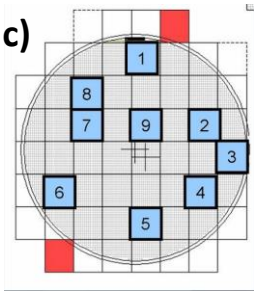


Figure 3.3 : a,b) Ségmentation d'une plaque en anneaux afin de mieux répartir les points de mesures c) Distribution des points de mesures par champ photolithographique.

Après avoir expliqué la méthode de mesure pour caractériser les différents paramètres physiques de la cellule eSTM, nous allons maintenant décrire les premiers résultats électriques qui nous ont poussés à effectuer la campagne de mesures.

3. Les résultats d'endurance

Les premiers tests électriques effectués sur la cellule eSTM consistent à exécuter un programme d'endurance de 500k cycles à 105°C. L'objectif de ce test est d'étudier la fiabilité et la robustesse de l'architecture mémoire. La Figure 3.4 représente la distribution des tensions de seuil V_T des points mémoires après le test d'endurance. Cette distribution manifeste des cellules extrinsèques qui n'ont pas été effacées correctement. Cependant, l'exclusion des cellules appartenant à la première et deuxième ligne de bit du test d'endurance montre une distribution de V_T tout à fait normale. Pour déterminer la source de ce problème d'effacement, une liste de paramètres physiques a été établie à l'aide d'une étude théorique de l'architecture eSTM et en prenant en compte l'expérience des ingénieurs de ST-Rousset.

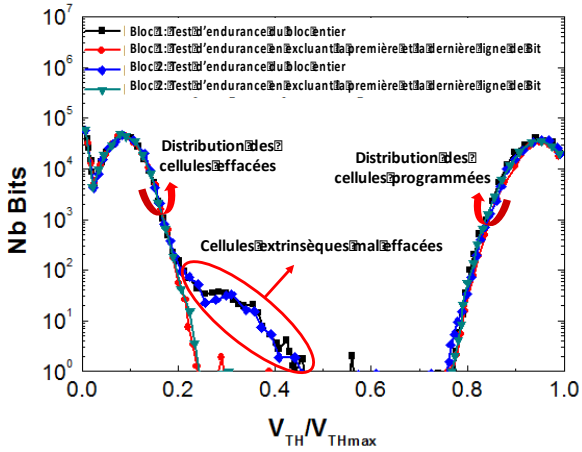


Figure 3.4 : Distribution du V_T après un cyclage de 500k à 105°C

La Figure 3.5 liste les différentes criticités de cette architecture ainsi que leurs localisations. La campagne de mesures consistera à étudier la variabilité de chaque paramètre et de les corréler avec les différents résultats électriques. Ces paramètres critiques sont quasiment identiques à ceux d'une mémoire Flash NOR. En plus des problématiques au niveau de la zone d'active, la grille flottante (Poly1) et la définition du transistor mémoire (Poly2) s'ajoutent aux problématiques liées au transistor vertical qui sont propres à l'eSTM. Ce transistor vertical met en avant quatre points critiques : la largeur de la tranchée, l'implant source, l'oxyde de grille ainsi que la distance entre le transistor vertical et les points mémoires.

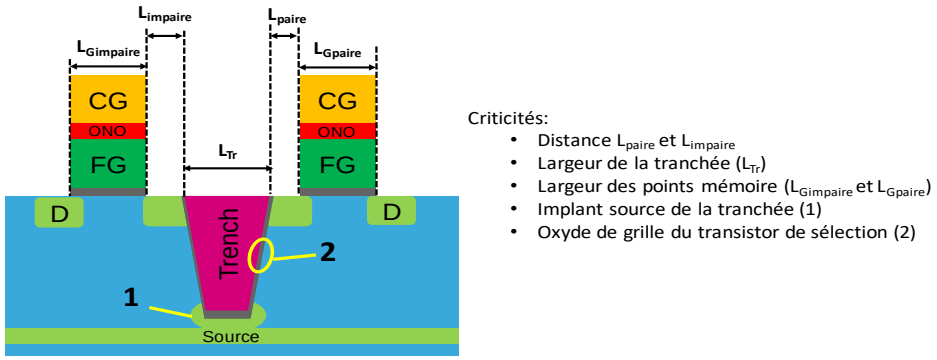


Figure 3.5 : Liste des différentes criticités de l'architecture eSTM ainsi que leurs localisations.

La prochaine partie va nous permettre de présenter en détail les résultats de l'analyse de variabilité des différents paramètres physiques de la cellule eSTM.

II. Analyse de variabilité : Résultats

Dans le paragraphe précédent nous avons mis la lumière sur les paramètres critiques de l'architecture eSTM. Une campagne de mesures et des expériences ont été effectuées pour vérifier chaque paramètre cité dans la Figure 3.5. Dans un premier temps nous allons évaluer la robustesse des procédés utilisés pour effectuer l'implantation des dopants de la source et la croissance de l'oxyde de grille du transistor vertical. Puis nous allons détailler les résultats des mesures en ligne au niveau de la zone active, du transistor vertical, de grille flottante et du transistor mémoire.

1. L'implant source au fond de la tranchée

L'implant source du transistor de sélection est réalisé après l'étape de gravure. Cette étape est critique, car l'implantation des dopants permet de connecter le futur transistor vertical à sa source en évitant d'implanter les zones aux alentours, comme les flancs de la tranchée ou les actives sous la couche du masque dure AHM. Pour ajuster les conditions d'implantation, des simulations TCAD ont été réalisées avec différentes énergies d'implantation pour l'arsenic (dopage type N pour la pointe de fond de la tranchée) et différentes doses de phosphore (dopage type N pour l'implant NISO). Les résultats de simulation sont regroupés en Figure 3.6. Les conditions qui permettent d'avoir une cellule fonctionnelle avec un compromis entre l'énergie d'implantation et la morphologie de la connexion sont entourées en rouge sur la Figure 3.10 (As : $5e^{13}cm^{-3}$40keV / P : $2e^{13}cm^{-3}$ 500keV).

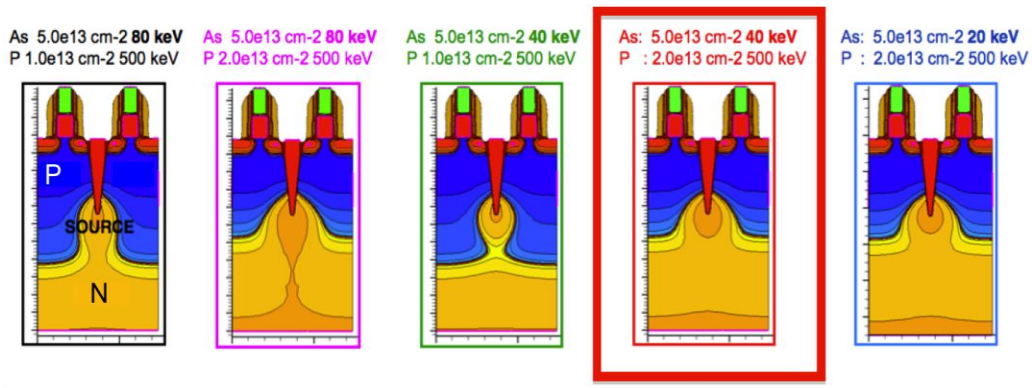


Figure 3.6 : Simulation TCAD de différents implants source et NISO. Les conditions d'implantation retenues sont entourées en rouge.

L'implant source étant défini, nous devons vérifier si la couche de protection AHM restante après la gravure permet de bien stopper l'implant choisi. Pour cela nous avons implanté plusieurs plaques avec différentes épaisseurs de AHM. L'effet d'un recuit sur le masque dur AHM a aussi été évalué. Le Tableau 3.1 résume les tests effectués :

Numéros de test	Épaisseur de AHM	Recuit	Implant As
1	2000Å	Non	Oui
2	1500Å	Non	Oui
3	1000Å	Non	Oui
4	1500Å	Non	Oui
5	1500Å	Oui	Oui

Tableau 3-1 : Tests d'arrêt de l'implant source pour différents cas de masque de protection

Les trois premières plaques vont permettre d'estimer l'impact de l'épaisseur de AHM sur l'implantation. La plaque 4 est un double de la plaque 2 et le AHM de la plaque 5 subira le recuit. Des analyses SIMS (Spectrométrie de Masse à Ionisation Secondaire) sur les zones actives ont été réalisées afin de vérifier si le silicium sous le AHM a bien été protégé. Ce type d'analyse permet de connaître la concentration et la profondeur de l'arsenic (ou d'autres dopants / impuretés) dans un matériau [69]. La Figure 3.7 représente la concentration d'arsenic en fonction de la profondeur dans le silicium. Les résultats de ces tests montrent que peu importe l'épaisseur de la couche AHM et la présence de l'étape de recuit, la concentration d'arsenic à la surface est trop élevée (entre 10^{18} et 10^{19} d'atomes/cm^3). Il faut donc trouver une autre solution pour stopper davantage l'implantation afin d'être inférieur à une valeur acceptable de 10^{17} atomes/cm^3 en surface.

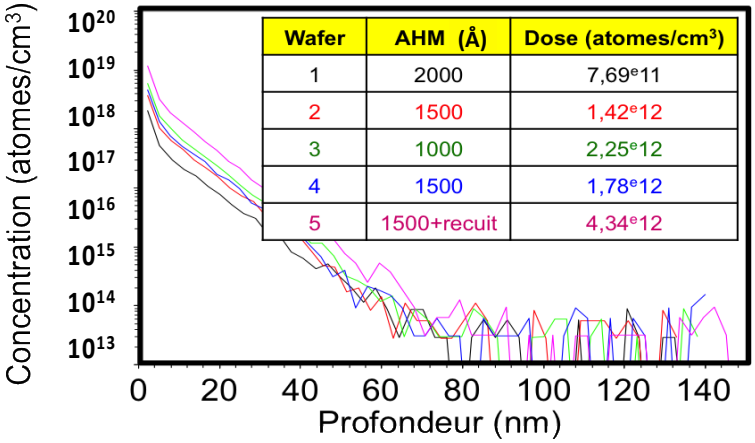


Figure 3.7 : Analyse SIMS indiquant la concentration de l'arsenic pour différents masques de protection à base de AHM lors de l'implantation de la source du transistor de sélection.

Afin de stopper davantage l'implantation, de nouveaux tests ont été réalisés en ajoutant une couche d'oxyde sacrificiel de 20nm au-dessus du silicium avant de déposer les différentes couches de AHM (Tableau 3.2).

Numéros de plaque	Oxyde sacrificiel	AHM	Implant As
1	Oui	2000Å	40keV
2	Oui	2000Å	20keV
3	Oui	2000Å	10keV
4	Oui	-	20keV

Tableau 3-2 : Tests d'arrêt de l'implant source pour différents cas de masque de protection avec la présence d'oxyde sacrificiel sous l'AHM

La Figure 3.8 représente les nouveaux résultats des analyses SIMS des plaques tests. La plaque de référence (N°4 en gris) composée de silicium et d'oxyde sacrificiel sans AHM ne stoppe pas les implants. Par contre, les trois autres courbes (plaques 1,2 et 3) arrêtent bien l'implant, peu importe l'énergie d'implantation (40keV, 20keV et 10keV) étant donné que la concentration à la surface du wafer n'excède pas les 10^{17} atomes/cm^3.

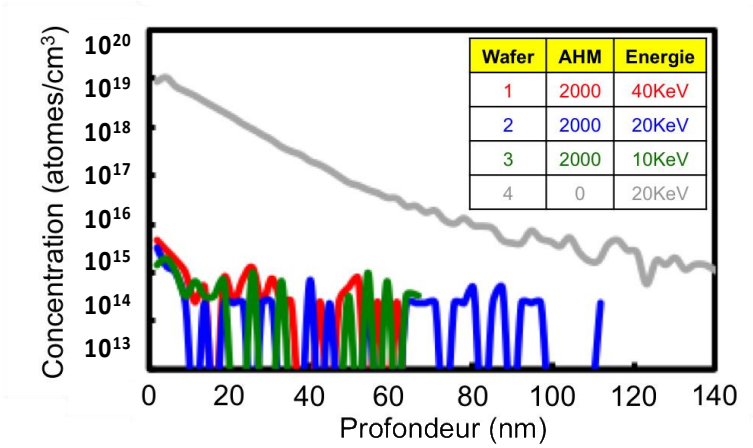


Figure 3.8 : Analyse SIMS indiquant la concentration d'atomes d'arsenic pour différents masques de protection à base de AHM et d'oxyde sacrificiel lors de l'implantation de la source du transistor de sélection

Pour éviter que l'implantation de la source du transistor de sélection ait un impact sur les performances de la partie logique du produit, il est important de coupler un oxyde sacrificiel avec un masque de AHM.

2. L'oxyde de grille du transistor de sélection

L'oxyde de grille d'un transistor représente un paramètre très important pour le bon fonctionnement de ce dernier. Cet oxyde doit être assez épais et uniforme tout au long de la tranchée afin de permettre d'éviter les fuites entre le transistor de sélection et le substrat. Il existe deux façons de faire croître cet oxyde :

- Une oxydation au four
- Une oxydation par ISSG

Avant de montrer les résultats des deux méthodologies, il convient de présenter ces procédés d'oxydation.

L'oxydation par ISSG est une oxydation humide par réaction chimique gazeuse (H_2, O_2, N_2) sur des wafers chauffés sous lampe UV, comme expliqué en Figure 3.9. Des sondes situées sous le wafer permettent de réguler la température dans la chambre afin de contrôler le processus. Un apport d'eau (sous forme de dihydrogène H_2 et dioxygène O_2) permet la réaction avec le silicium afin de créer de l'oxyde (SiO_2) : $2H_2O + Si => SiO_2 + 2H_2$. Cette oxydation permet de monter rapidement en température et de mieux contrôler la croissance de l'oxyde, ce qui rend l'oxydation plus uniforme. L'oxyde formé par ISSG présente généralement un courant de fuite particulièrement faible comparé aux autres méthodes de croissance / dépôt d'oxyde. Son inconvénient réside dans son impossibilité à obtenir des oxydes épais [54], [70].

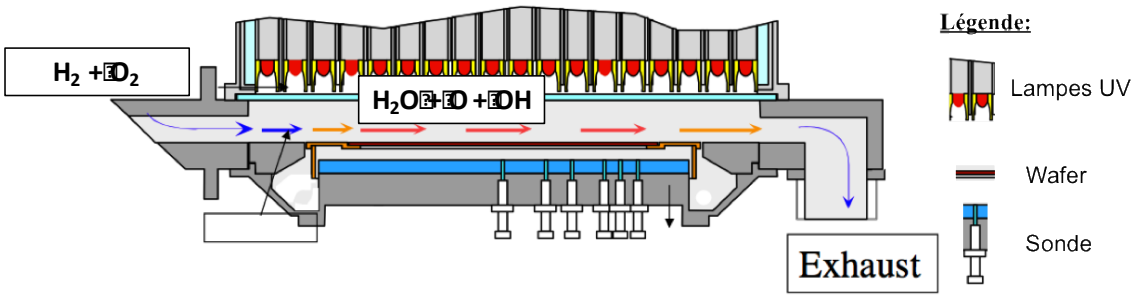


Figure 3.9 : Principe de l'oxydation ISSG [Interne ST]

L'oxydation en four est basée sur le même principe que l'oxydation par ISSG, c'est-à-dire l'injection de gaz H_2 et O_2. Comme son nom l'indique, cette oxydation se passe dans un four où l'atmosphère interne est chauffée, comme représentée sur la Figure 3.10. Contrairement au processus avec les lampes UV, le four possède une inertie plus importante, mais permet d'obtenir des épaisseurs d'oxyde importantes.

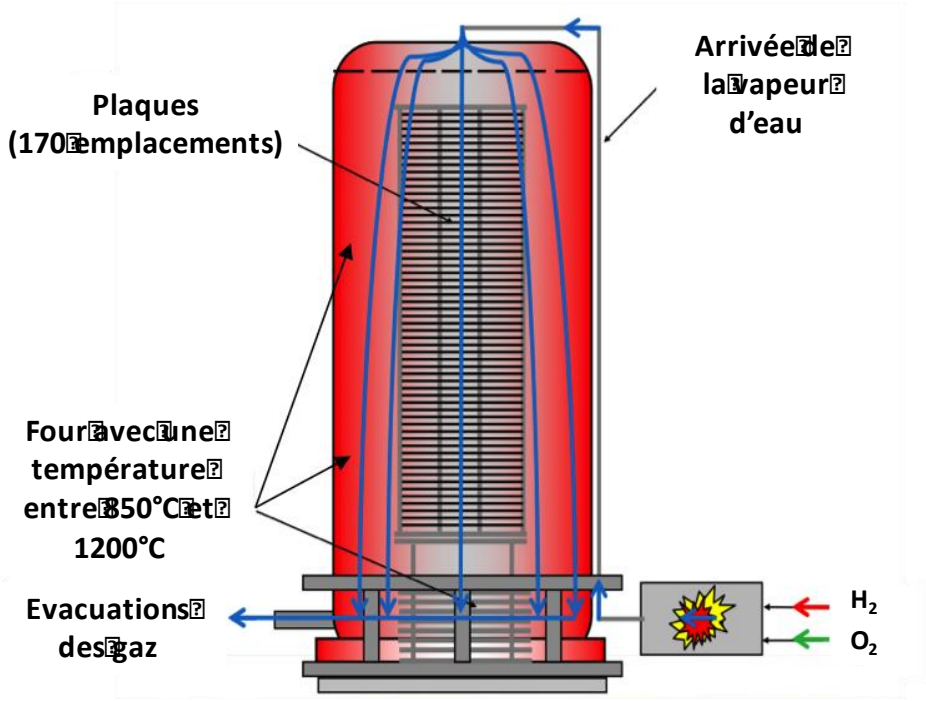


Figure 3.10 : Principe de l'oxydation par four [Interne ST]

Grâce à des simulations TCAD, l'épaisseur d'oxyde a été initialement fixée à 7.5nm. La topologie de l'oxyde dans la tranchée de l'eSTM a été comparée pour les deux types d'oxydation (Figure 3.11). Les deux types d'oxydation ont des comportements différents dans la tranchée. La Figure 3.11a montre que les résultats de l'oxydation en four ne sont pas uniformes avec une différence de 5nm entre les flancs et les angles de la tranchée. En revanche la Figure 3.11b montre que l'oxydation par ISSG est beaucoup plus uniforme avec seulement une variation de 1nm entre les flancs et les angles de la tranchée.

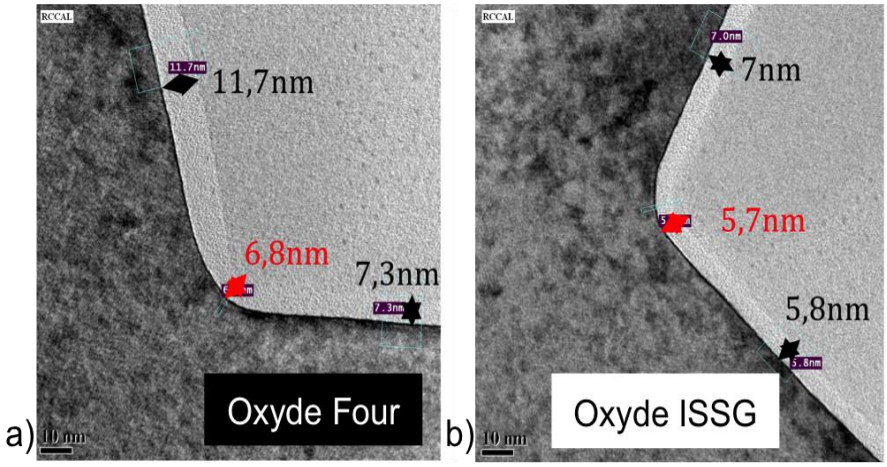


Figure 3.11 : Résultats d'oxydation a) par four et b) par ISSG pour une épaisseur de 7.5nm dans la tranchée de la cellule eSTM [Interne ST]

Afin de garder une bonne uniformité et un bon contrôle d'épaisseur d'oxyde, la méthode ISSG a été retenue. Une bonne uniformité de l'oxyde de grille du transistor de sélection permettra de garantir les performances électriques de ce transistor durant les tests de fiabilité.

3. L'analyse de variabilité des lignes de la zone active

Les mesures électriques de la Figure 3.4 ont souligné un dysfonctionnement des cellules appartenant à la première et deuxième ligne de bit. Cette anomalie nous a orienté vers une mesure intra-die de la largeur des lignes de la zone active. Cette étude de variabilité intra-die consistera à comparer les lignes d'active au bord du plan mémoire et celles au milieu du plan mémoire. La Figure 3.12 met en avant la localisation des points de mesures effectuées dans le plan mémoire.

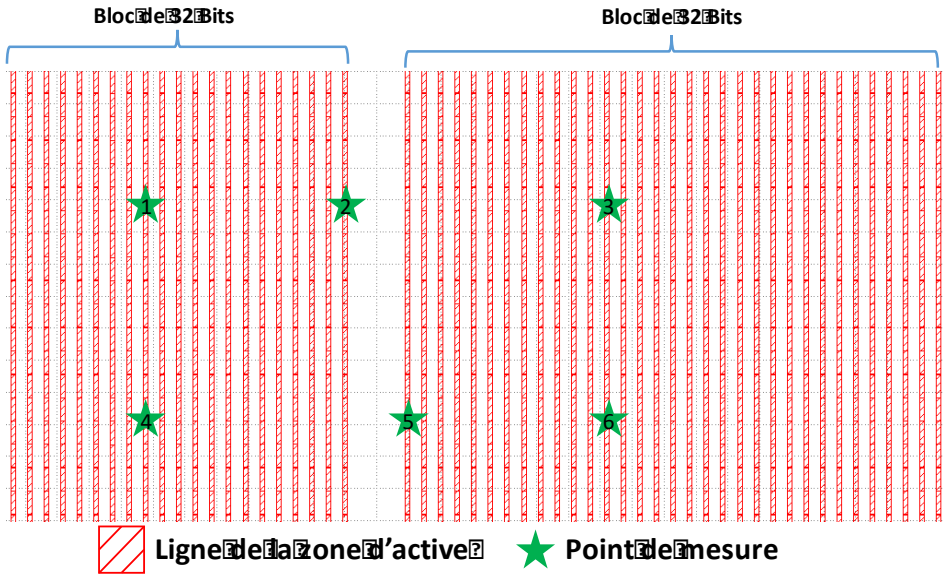


Figure 3.12 : Localisation des points de mesures intra-die des lignes de la zone d'active effectuées dans le plan mémoire

Dans ce cas d'étude, il faudra plus d'une heure pour mesurer tous les champs d'une plaque (162 points de mesures). C'est pour cette raison que neuf sites de mesures ont été requis

pour cette étude. La Figure 3.13 montre que les largeurs des lignes de la zone active au bord du plan mémoire sont plus importantes que celles au milieu du plan mémoire. La dispersion après gravure est due principalement à l'apparition d'une pente plus importante sur le bord du plan mémoire (Figure 3.14a).

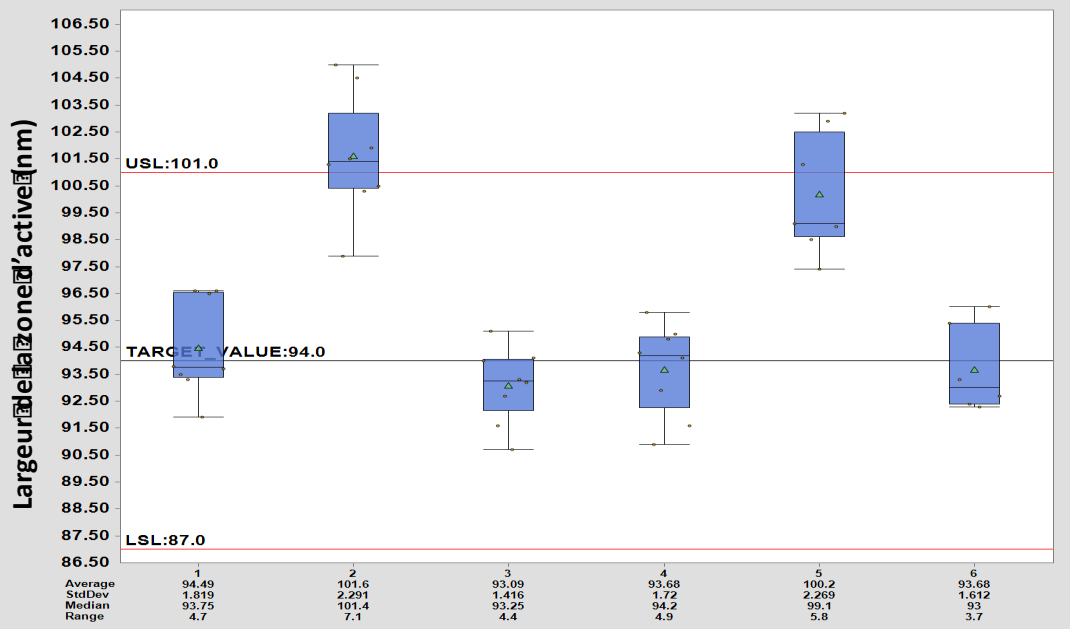


Figure 3.13 : Mesure intra-die de la largeur de zone d'active après gravure

Cette dispersion intra-die corrèle bien avec le problème d'effacement des cellules appartenant à la première ligne de bit dans le test d'endurance présenté précédemment. C'est pour cette raison qu'un redimensionnement des lignes d'active au bord du plan mémoire a été effectué sur un nouveau réticule en utilisant une Correction Optique de Proximité (*Optical Proximity Correction* : OPC). La Figure 3.15 montre que cette correction permet d'assurer des lignes d'actives de largeur identique, quelle que soit la position dans le plan mémoire.

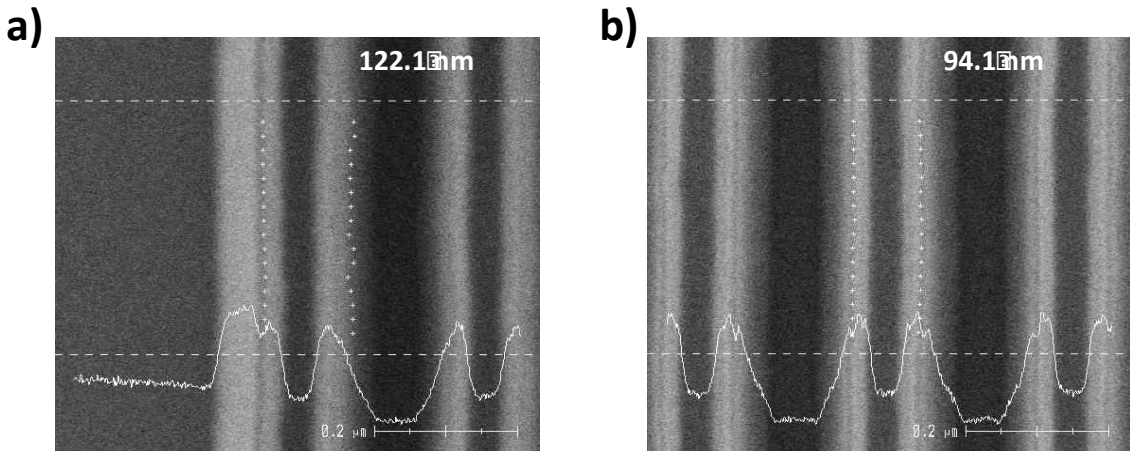


Figure 3.14 : Largeur de la zone d'active après gravure a) bord du plan mémoire b) milieu du plan mémoire – La pente de la zone d'active au centre du plan mémoire est plus abrupte que celle au bord du plan mémoire.

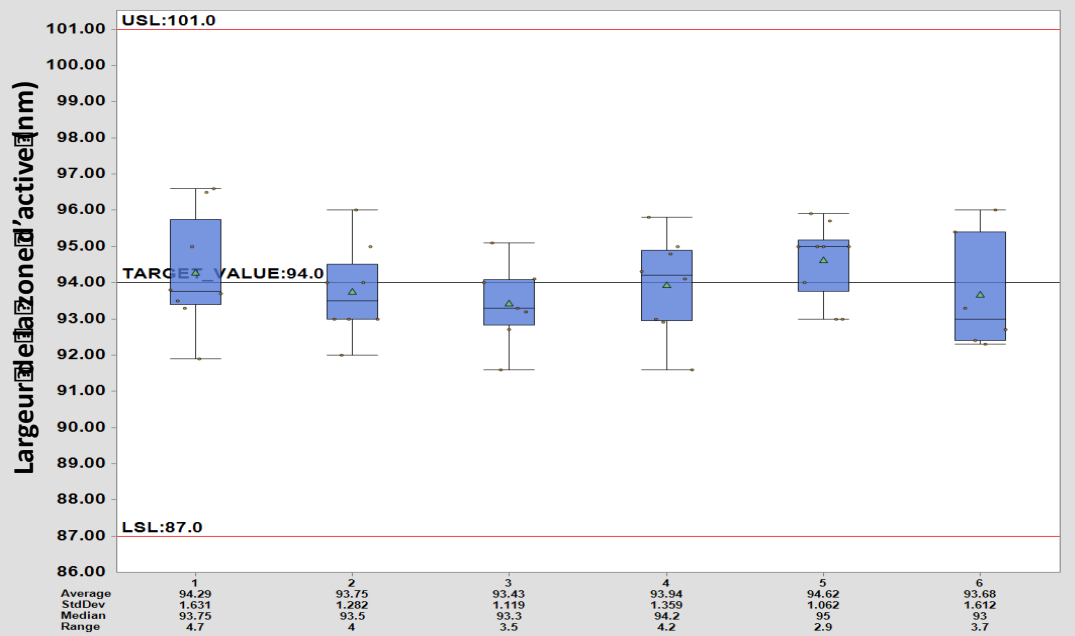


Figure 3.15 : Mesure de la dispersion intra-die de la zone d'active après la correction optique de proxémité

Après avoir étudié la dispersion intra-die, l'objectif est l'étude de la variabilité intra-plaque de la ligne d'active après gravure. La Figure 3.16a montre la cartographie de la mesure effectuée ainsi que la dispersion en fonction du rayon de la plaque (Figure 3.16b). La dispersion observée est aux alentours de 5nm en moyenne. Ce résultat a été observé sur plusieurs plaques provenant de lots différents. Cette dispersion est tout à fait normale ; elle ne représente que 5% de la valeur demandée (95nm).

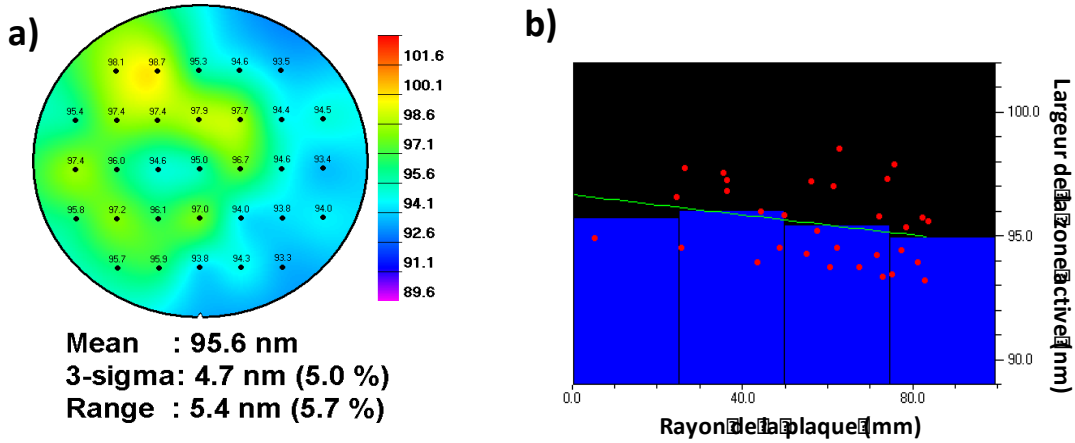


Figure 3.16 : a) Cartographie de la mesure intra-plaque de la zone active b) Représentation en fonction du rayon de la plaque

4. La grille flottante

La mesure effectuée au niveau de la grille flottante correspond à l'espace entre deux lignes de polysilicium (Figure 3.17). Le choix de ce paramètre permet d'évaluer les effets de couplages électrostatiques entre deux cellules mémoires. Ces effets sont plus importants pour les nœuds technologiques les plus avancés (28nm) que pour l'eSTM (80nm).

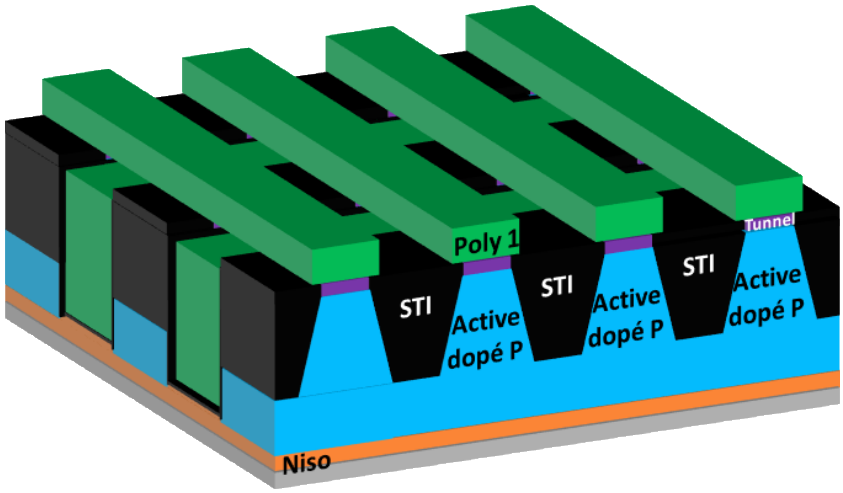


Figure 3.17 : Représentation de la cellule eSTM après gravure de la grille flottante

Dans ce paragraphe l'analyse de variabilité se déroulera en deux parties, tout d'abord une analyse de la variabilité intra-champs. Cette étude consistera à mesurer les différentes puces appartenant à un seul champ de photolithographie. Enfin les résultats de la variabilité intra-wafer seront exposés.

Une plaque du produit eSTM comporte 27 champs de photolithographie et chaque champ comporte 13 puces fonctionnelles (Figure 3.18). Pour limiter le temps de mesure, nous avons opté pour une mesure de trois puces par champ comme indiqué sur la Figure 3.18 avec des étoiles en noir. Ce choix nous permettra de mesurer tous les champs pour évaluer en même temps la variabilité intra-champ et intra-wafer.

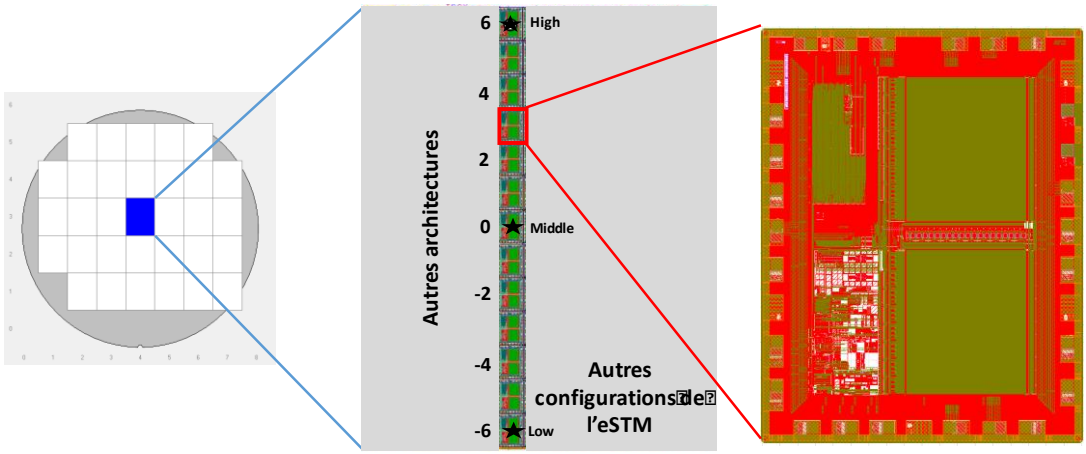


Figure 3.18 : Représentation de la puce eSTM dans un champ photolithographie

Pour représenter les mesures après gravure, nous avons opté pour une cartographie de chaque puce mesurée (Figure 3.19). Nous constatons dans un premier temps que la dispersion intra-wafer est identique pour les trois points de mesures. Cette dispersion de 6nm reste normale par rapport à d'autres produits de STMicroelectronics-Rousset. Cette campagne de mesures a mis la lumière sur une dispersion intra-champ de 10nm. Cette dispersion est anormalement élevée et aucun test électrique n'a montré des effets liés à ce problème.

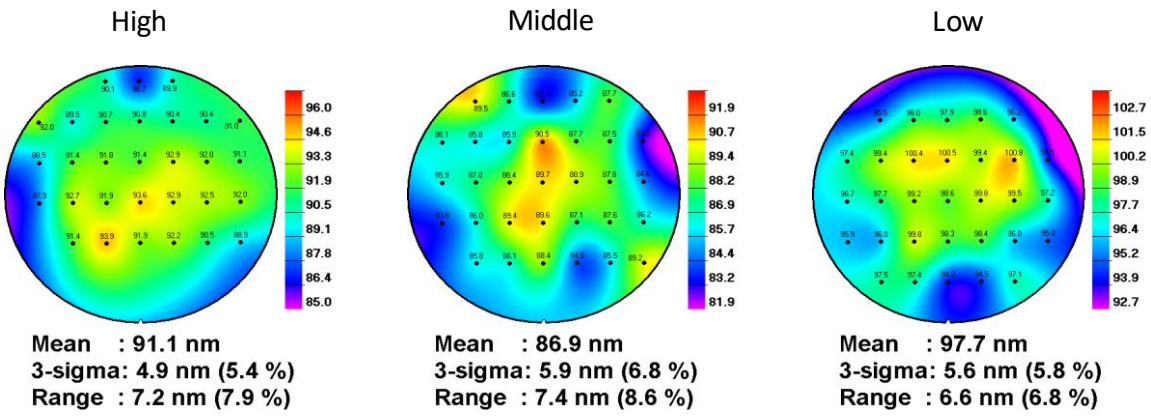


Figure 3.19 : Cartographie de la mesure intra-champ de la grille flottante après l'étape de gravure

Dans le but de trouver la source de cette variabilité intra-champ nous avons commencé par vérifier l'étape de photolithographie ainsi que le réticule utilisé. Le nombre de points de mesures par champ photolithographie a été augmenté pour une meilleure caractérisation de la variabilité intra-champ. Les résultats sont présentés sur la Figure 3.20. Il apparaît que la dispersion intra-champ est aussi importante qu'au niveau de la mesure après gravure. Cependant, la mesure sur le réticule ne montre aucune dispersion.

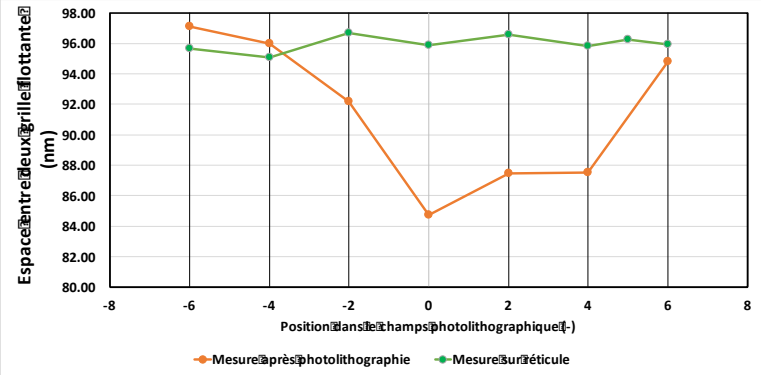


Figure 3.20 : Mesure de la variabilité intra-champ de l'espace entre deux grilles flottantes après photolithographie

La dispersion de 10nm trouvée durant la mesure après photolithographie (Figure 3.20) identifie le procédé de photolithographie comme responsable de la variation intra-champ. Cependant après vérification les ingénieurs de l'atelier photolithographie estiment que cela est dû aux imperfections du réticule, et plus précisément à la planéité de ce dernier. La planéité du réticule est définie comme étant l'écart admissible de la surface chrome à partir d'un plan de référence défini. Les réticules sont classifiés comme ci-dessous :

- 0.5T : Ces réticules ont une planéité inférieure ou égale à 0.5µm, ce qui signifie que la surface chromée du réticule peut être alignée entre deux plans parallèles qui sont à 0.5µm d'intervalle
- 1T : Ces réticules ont une planéité inférieure ou égale à 1µm
- 2T : Ces réticules ont une planéité inférieure ou égale à 2µm

Ces erreurs au niveau de la planéité de réticule induisent une déviation du plan d'image (*Image Plan Deviation* : IPD) au niveau du wafer. Cette IPD est estimée à 1/20 de la valeur de la planéité réelle du réticule. En conséquence, l'écart de mise au point induite par une certaine forme de réticule (Figure 3.21) sera au niveau du wafer de l'ordre :

- Un maximum de 25nm pour un réticule 0.5T
- Un maximum de 50nm pour un réticule 1T
- Un maximum de 100nm pour un réticule 2T

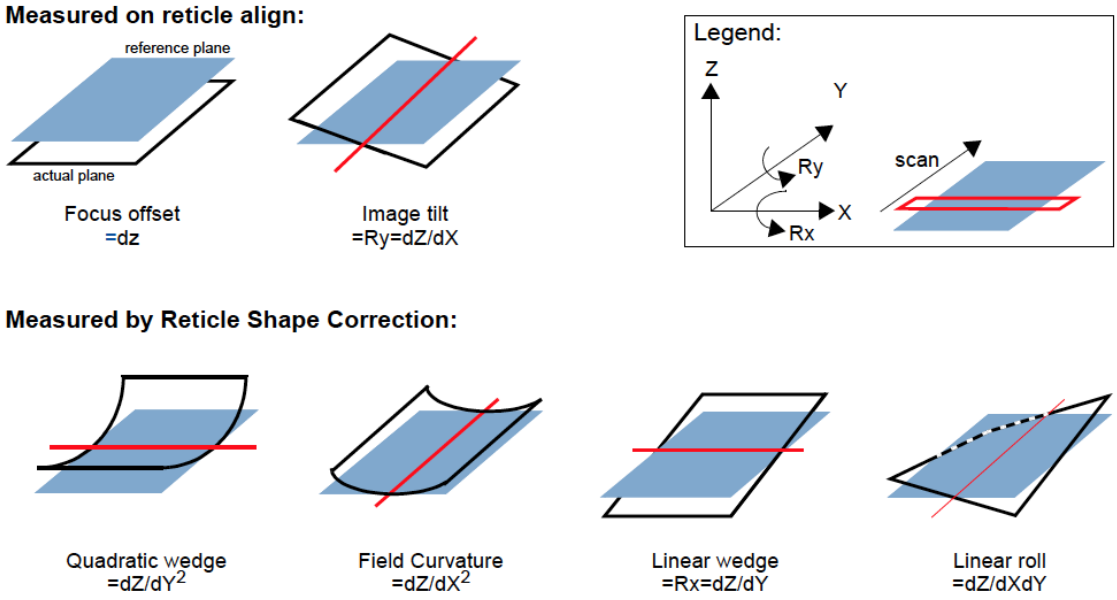


Figure 3.21 : Les différentes formes de déformation de réticule (Source : ASML)

Dans notre cas la planéité du réticule utilisé est de 1T, ce qui peut expliquer la variation intra-champ de 10nm trouvée durant cette étude. En théorie, il est possible de corriger ces imperfections en activant l'option *Reticle Shape Correction* (RSC) dans l'équipement de photolithographie [71]. Cette option permettrait d'utiliser des réticules moins plats (moins chers) et de maintenir un contrôle de mise au point standard. Cependant la mise en place du RSC n'a pas pu être effectuée, car cette manipulation nécessite l'arrêt de l'équipement durant plusieurs jours. Le fait que cette variation soit comprise dans les limites de spécifications et n'impacte pas le fonctionnement de la cellule ne justifie pas l'arrêt d'un équipement de production.

5. Variabilité liée au transistor de sélection

Le développement de la cellule eSTM est basé sur une bonne intégration du transistor vertical. Une étude approfondie de ce transistor a été requise afin de mieux comprendre le comportement de ce nouveau dispositif. Comme expliqué dans le chapitre 1 un masque de protection est primordial pour protéger les zones à ne pas graver durant la réalisation de la tranchée (Figure 3.22). Pour le besoin de l'application finale du produit, la largeur de la tranchée ciblée en photolithographie (88nm) est à la limite des outils de production présents sur le site de ST-Rousset. Dans le but d'atteindre cette résolution, une résine sensible à la longueur d'onde 193nm a été utilisée.

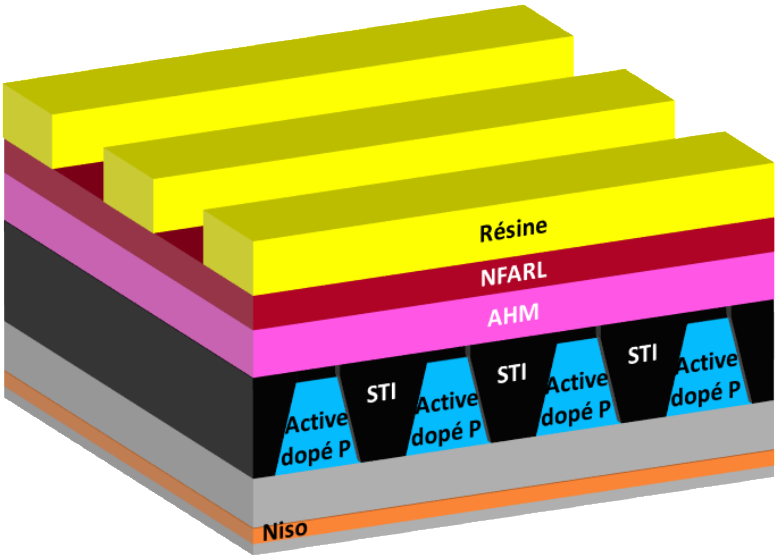


Figure 3.22 : Détail des couches de protection avant gravure de la tranchée

Le fait de vouloir atteindre une résolution proche de la limite des équipements de photolithographie augmente la probabilité d'induire une variabilité du procédé dans le temps. Dans cette optique, les résultats de l'étude de l'impact de cette contrainte technologique sur la variabilité intra-die seront présentés suivis des résultats de la variabilité temporelle.

a) La variabilité intra-die

L'analyse de la variabilité intra-die a été motivée par le besoin de caractériser cette nouvelle étape de procédé de fabrication qui est une première pour STMicroelectronics-Rousset. Ces variations intra-die peuvent être classées en deux catégories : systématiques et aléatoires. Les variations systématiques ont un degré de corrélation entre les différents dispositifs qui changent avec la distance et qui sont liées à des imperfections de fabrication. De l'autre côté les variations aléatoires sont complètement indépendantes du layout du dispositif. Elles sont dues la plupart du temps aux limites de la miniaturisation.

La Figure 3.23a montre une capture d'écran du *Graphic Database System* (GDS) correspondant au plan mémoire sur lequel les mesures ont été effectuées. Dans ce cas d'étude, les sites de mesures choisis (Figure 3.23b) ont pour objectif de caractériser les effets liés au *micro-loading*. Ce phénomène peut affecter l'uniformité de la largeur des tranchées entre celles appartenant à un environnement dense et celles appartenant au bord du plan mémoire.

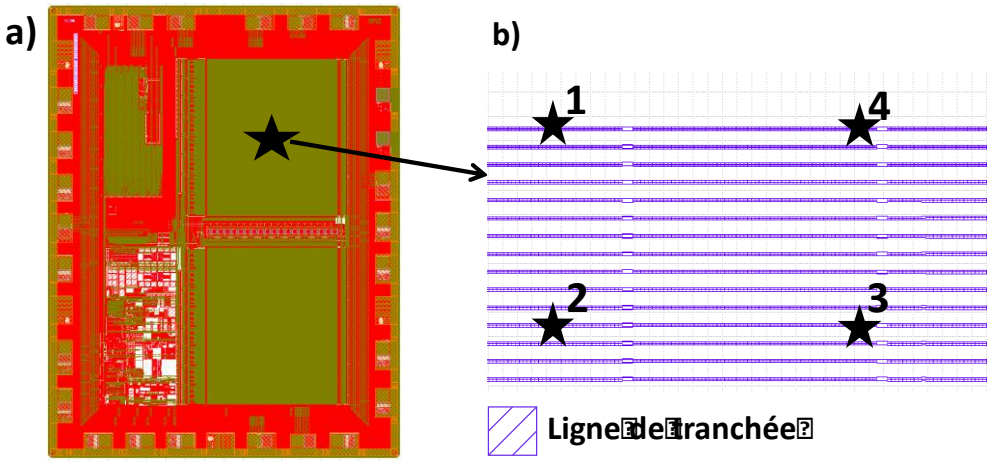


Figure 3.23 : a) Capture d'écran du GDS correspondant au plan mémoire b) Localisation des points de mesure intra-die

La Figure 3.24 confirme l'existence du phénomène du *micro-loading* [72]. La dispersion due à ce phénomène est de l'ordre de 24nm en moyenne. Une telle variation ne peut être tolérée au risque d'avoir un impact sur le bon fonctionnement de la cellule mémoire. Cependant l'évaluation de cet impact au niveau électrique n'a pas pu être réalisée, car le programme de test avait été modifié pour éliminer les tranchées au bord du plan mémoire pour éviter de retarder les plans de qualifications de la technologie suite aux observations de la Figure 3.24.

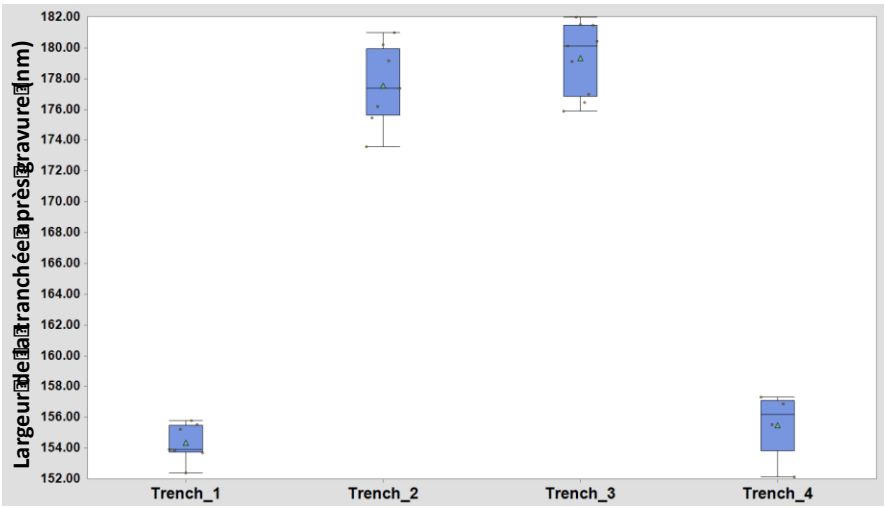


Figure 3.24 : Mesure intra-die de la tranchée après gravure

Cet effet de *micro-loading* peut être évité en appliquant une modification au niveau de la conception du plan mémoire. La première solution consiste à utiliser une Correction Optique de Proximité (OPC) pour modifier la largeur de la tranchée en bord du plan mémoire [73]. La deuxième solution implique l'ajout de tranchées non connectées au plan mémoire pour simuler un environnement dense. Cependant la mise en place de l'OPC nécessite un développement de trois mois, et par manque de temps et de ressources le choix des tranchées non connectées a été le plus judicieux.

b) La variabilité temporelle

La variabilité temporelle provient d'une variation de la matière première ou d'une dérive des procédés de fabrication [74]. Comme expliqué auparavant, le besoin d'atteindre une dimension à la limite de la capacité des équipements de photolithographie engendrera une dispersion dans le temps (Figure 3.25). L'objectif de cette partie est de comprendre l'impact de cette variabilité sur l'architecture et le fonctionnement de l'eSTM.

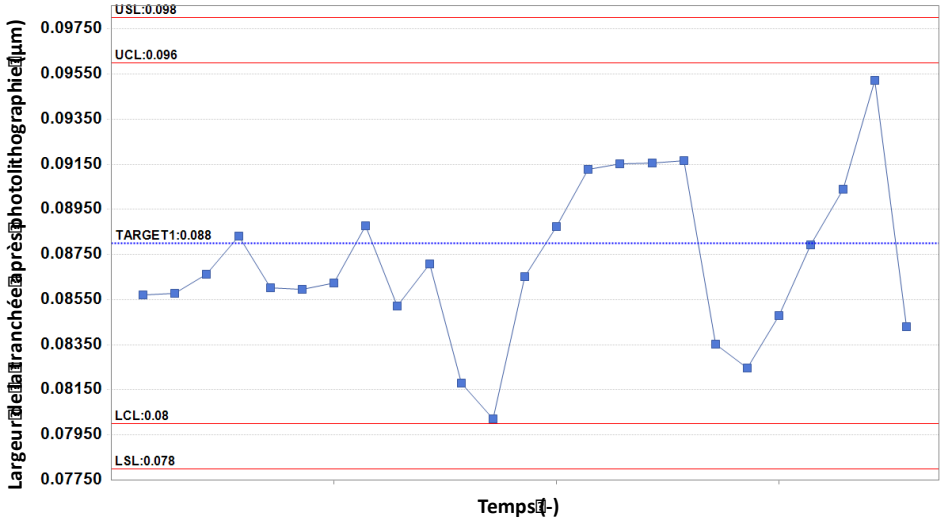


Figure 3.25 : La variation temporelle de la largeur de la tranchée après photolithographie

La variabilité du procédé de fabrication observée précédemment est due à la tolérance de l'équipement de photolithographie et aussi au procédé de la définition du transistor de sélection. La Figure 3.26 résume les différents scénarii issus de ces problématiques.

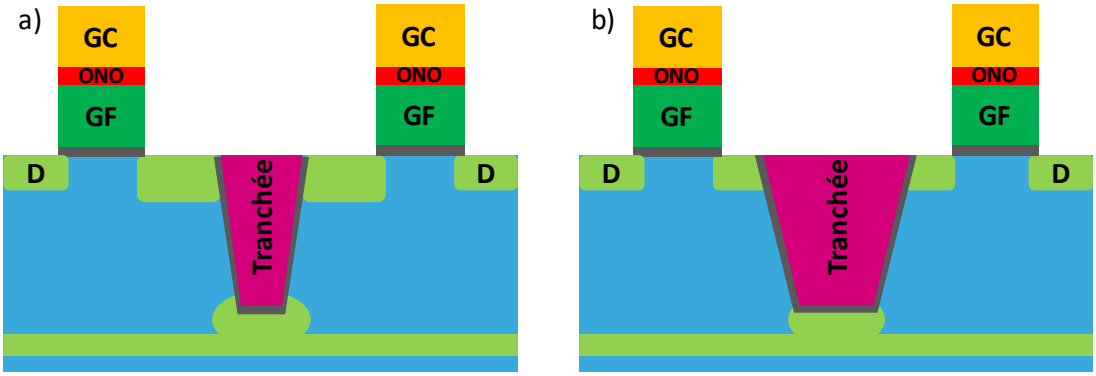


Figure 3.26 : L'impact de la largeur de la tranchée sur l'implant entre la tranchée et le transistor mémoire

La variation de la largeur de tranchée représentée dans la Figure 3.25 va avoir un impact sur la distance entre le transistor de sélection et le transistor mémoire, ce qui va engendrer une modification de la profondeur de l'implant situé entre ces deux transistors. Les cellules représentées sur la Figure 3.26a sont plus éloignées du transistor de sélection ce qui laisse apparaître un implant plus profond que celui de la Figure 3.26b. Cette variation génère une différence sur la longueur effective (L_{eff}) du transistor de sélection qui va avoir pour effet de modifier la quantité de charges injectées dans les cellules de la Figure 3.26a durant la phase de programmation.

Dans le but de démontrer l'analyse théorique faite ci-dessus, un lot d'expériences avec des largeurs de tranchée (L_{Tr}) différentes au niveau de la gravure a été établi. Ce lot a été réalisé dans le cadre de la thèse de Maria Rizquez [75], le Tableau 3.3 montre les résultats de mesure de L_{Tr} obtenus après gravure sachant que la largeur du cahier de charge est de 175nm.

Wafer ID	L_{Tr} (nm)	Wafer ID	L_{Tr} (nm)	Wafer ID	L_{Tr} (nm)
1	172.3	8	178	15	185.5
2	172.6	9	167.4	16	161.5
3	175.3	10	164.9	17	184
4	178.7	11	187.4	18	185.5
5	186.9	12	173.9	19	161.3
6	177.6	13	175.1	20	162.1
7	172	14	169.4	21	162.8

Tableau 3-3 : Résultats des essais

Comme expliqué dans la partie précédente la distance entre le transistor mémoire et le transistor vertical est soupçonnée d'avoir un impact sur l'efficacité de programmation de la cellule ainsi que sur la tension de seuil effacée V_{TE}. L'analyse paramétrique (Figure 3.27) montre que la tension de seuil programmée V_{TP} augmente avec la largeur de la tranchée, alors que la tension de seuil effacée V_{TE} est moins dispersée lorsqu'il y a un amincissement de la tranchée (Figure 3.28). Cependant il faudrait tout de même rappeler que le test paramétrique est effectué sur des structures de test logées dans le chemin de la découpe (*scribe line*). Dans notre cas, contenu des contraintes de développement de l'eSTM comme technologie en développement, il est primordial d'effectuer les tests électriques directement sur le produit.

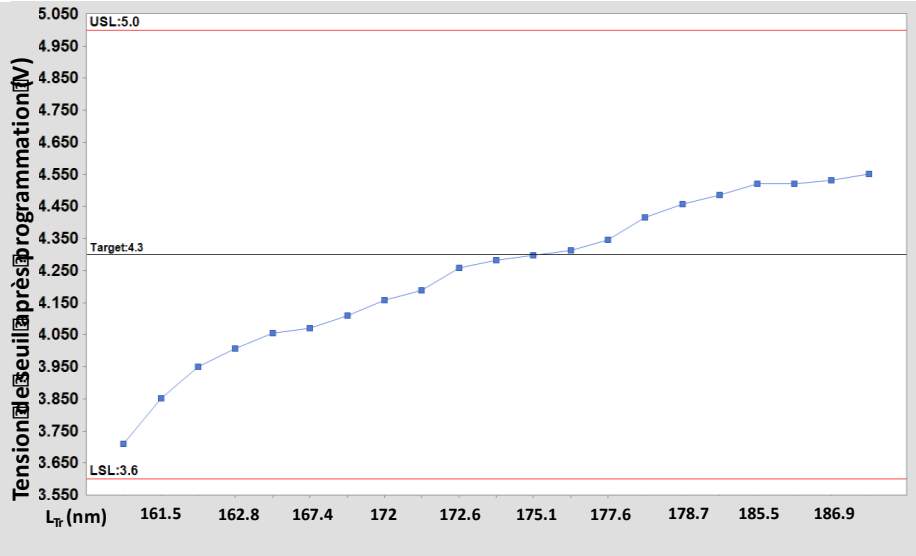


Figure 3.27 : Distribution de la tension de seuil V_{TP} en fonction de largeur de la tranchée

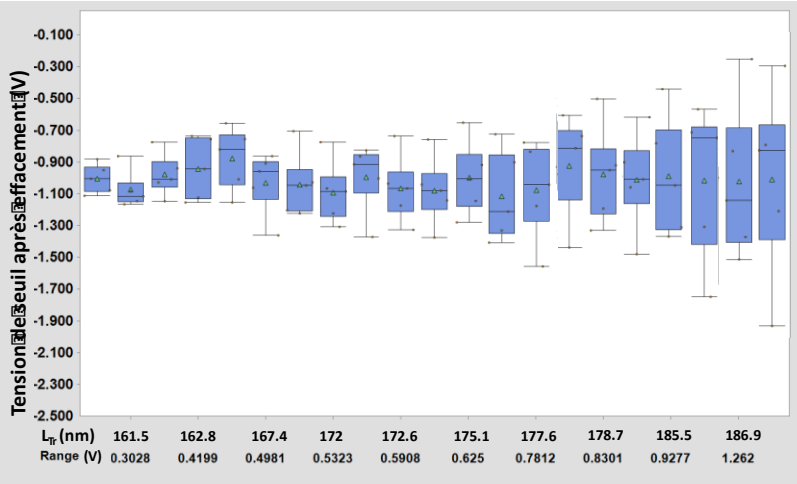


Figure 3.28 : Représentation de la dispersion de la tension de seuil V_{TE} en fonction de largeur de la tranchée

Dans le but d'évaluer l'impact d'une probable variation de la largeur de la tranchée sur le rendement du produit, un test EWS (Electrical Wafer Sort) a été effectué après le test paramétrique. Il s'agit d'une batterie de mesures permettant d'identifier les puces fonctionnelles et les puces défectueuses. Dans ce cas d'étude, nous allons nous concentrer sur les puces dont le disfonctionnement provient du plan mémoire et non pas à la périphérie (la partie logique). Une plaque du produit eSTM contient 442 puces et le rendement est calculé en utilisant l'équation ci-dessous :

$$Yield = \frac{Nombre\ de\ puces\ fonctionnelles}{Nombre\ total\ de\ puces} \times 100 \quad (\%)$$

Pour analyser les résultats du test EWS nous nous sommes intéressés au test appelé HB60, qui consiste à mesurer la fenêtre de programmation après un test d'endurance de mille cycles. Si la puce ne respecte pas le cahier des charges concernant la dégradation de la fenêtre de programmation, cette puce est alors considérée comme défectueuse. La Figure 3.29 montre que le nombre de puces défectueuses est proportionnel à la largeur du transistor de sélection. La dispersion de la largeur de tranchée a donc un grand impact sur le bon fonctionnement de la cellule ainsi que sur le rendement final.

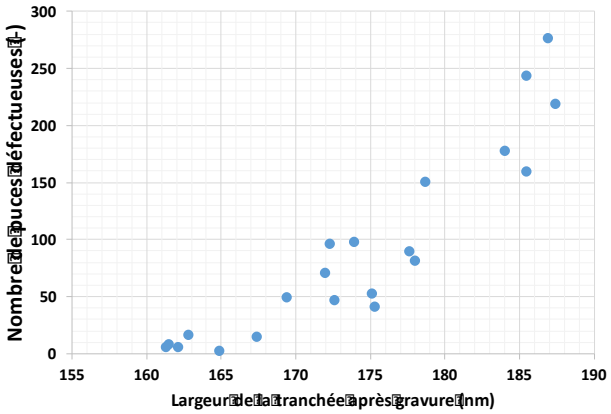


Figure 3.29 : L'impact de la largeur de la tranchée sur le nombre de puces défectueuses

Afin de mieux contrôler le procédé de fabrication de l'eSTM et de limiter l'impact de la distance entre la tranchée et le point mémoire, la mise en place d'une boucle de régulation a été proposée entre les deux opérations de gravure. La conception de cette boucle de régulation dépend de la robustesse de l'étape de gravure du transistor mémoire, dont l'étude de variabilité va être présentée dans la partie suivante.

6. La variabilité de la largeur de grille mémoire

Les variations intra-die systématiques ou aléatoires peuvent provoquer des différences dans les performances électriques de deux dispositifs identiques (même géométrie, layout et proximité). À l'échelle actuelle, il existe trois principales sources de variabilité intrinsèque pouvant affecter les performances électriques [76]:

- *Random Dopant Fluctuations* (RDF) : fluctuation aléatoire des dopants
- *Line Width Roughness* (LWR) : rugosité de ligne
- *Oxide Thickness Variations* (OTV) : variation de l'épaisseur d'oxyde

La campagne de mesure en ligne a montré l'existence d'une variabilité au niveau de la largeur de la grille mémoire qui est due à sa rugosité. La Figure 3.30 montre que la grille mémoire (Word Line) présente un élargissement au niveau de la première et la deuxième ligne de bit de l'ordre de 3nm.

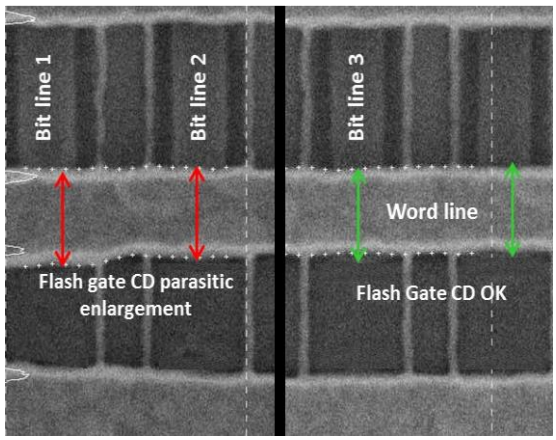


Figure 3.30 : Image SEMCD du plan mémoire [77]

La Figure 3.31 rappelle le premier test d'endurance (Figure 3.4) qui nous a amenés à faire cette campagne de mesure. Elle montre que la distribution de la tension de seuil V_T des points mémoires après le test d'endurance du bloc mémoire entier représente des cellules extrinsèques qui n'ont pas été effacées correctement. Cependant, l'exclusion des cellules appartenant à la première et deuxième ligne de bit du test d'endurance montre une distribution de V_T tout à fait normale. Cette batterie de mesures en ligne ainsi que les résultats du test paramétrique exposés dans le paragraphe II.5b montrent que cet élargissement de la grille mémoire peut être responsable des difficultés rencontrées lors de l'effacement des cellules.

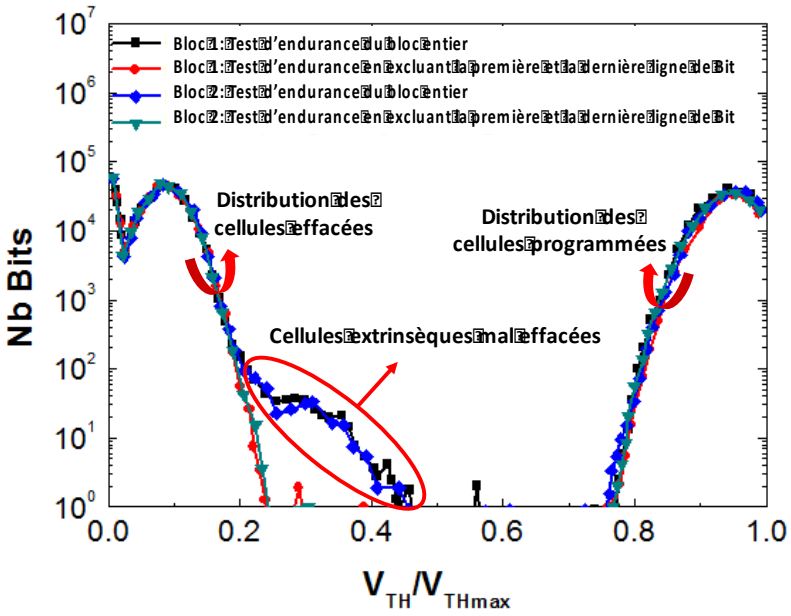


Figure 3.31 : Distribution du V_T après un cyclage de 500k à 105°C

Dans le but de mieux comprendre l'impact de cette rugosité sur les performances électriques, nous avons caractérisé plusieurs cellules avec différentes distances entre le transistor de sélection et le transistor à grille flottante. Les cellules utilisées pour ce test ont été obtenues en effectuant un désalignement volontaire entre le réticule de tranchée et celui de la grille mémoire. La Figure 3.32 présente un cas de figure où la cellule impaire est la plus éloignée du transistor de sélection.

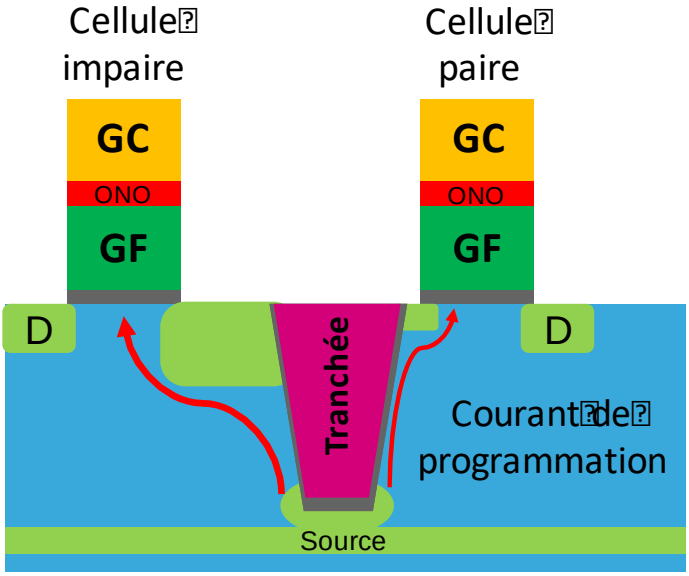


Figure 3.32 : Schéma d'une cellule désalignée

La caractérisation de cette cellule va permettre de comprendre l'impact de la distance entre le transistor de sélection et le point mémoire sur l'efficacité de programmation et la dégradation lors des tests d'endurance. Pour cela, un test d'endurance d'un million de cycles à température ambiante a été effectué. La Figure 3.33a représente la tension de seuil effacée et programmée en fonction du nombre de cycles. La tension V_{TP} de la cellule paire après le premier cycle est plus importante que celle de la cellule impaire. Ceci s'explique par la proximité de la

cellule paire de la zone de génération des porteurs chauds. La Figure 3.33a montre aussi une augmentation du V_{TE} de la cellule paire durant le cyclage. Pour expliquer cette augmentation nous avons effectué des courbes $I_D(V_G)$ de l'état effacé des deux cellules (Figure 3.33b) avant et après endurance. La pente sous le seuil de la cellule impaire reste intacte après le cyclage, par contre celle de la cellule paire se dégrade. Cette dégradation de la pente sous le seuil met en évidence un nombre important de défauts à la surface du canal du transistor mémoire. Dans le cas de la cellule impaire, l'implant plus important limite l'apparition de ces défauts comme illustrés sur la Figure 3.34.

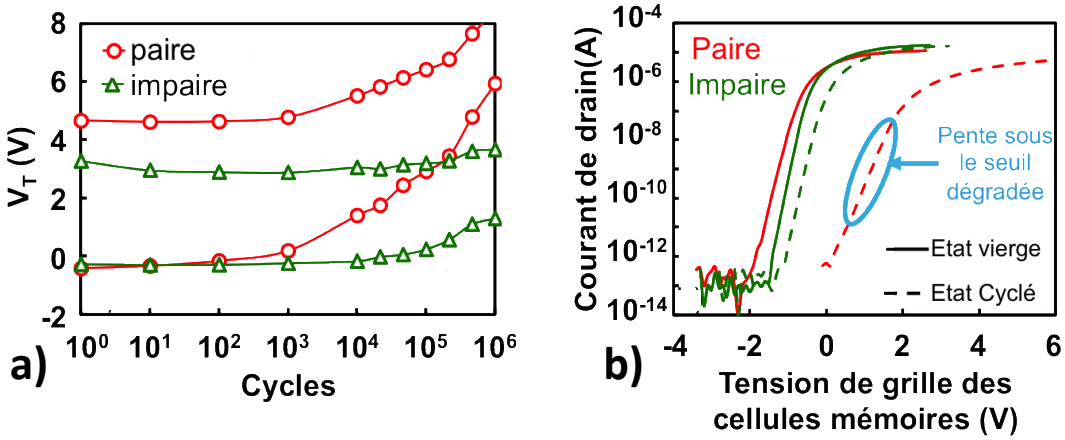


Figure 3.33 : a) Courbe de cyclage et b) $I_D(V_G)$ pour deux cellules eSTM dissymétriques avant et après cyclage

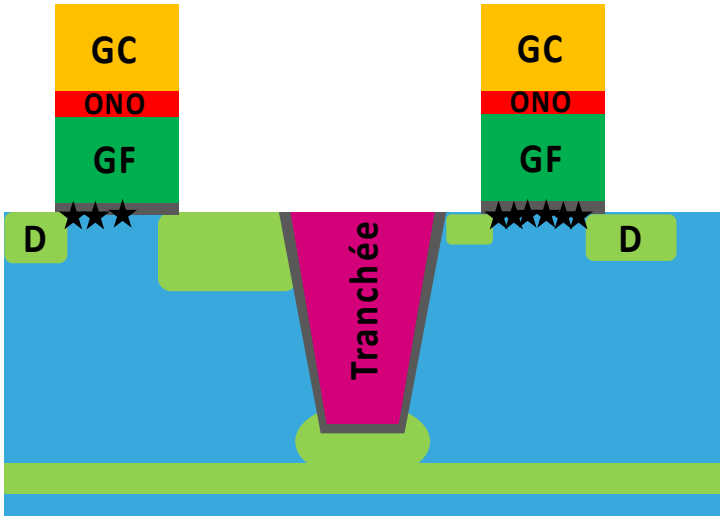


Figure 3.34 : Localisation de défauts après un test d'endurance pour une cellule eSTM désalignée

En analysant de plus près les résultats de la campagne de mesures effectuée au niveau de la grille mémoire, il s'est avéré que la rugosité touche tout le plan mémoire. Dans le but de quantifier cette rugosité, des mesures sur la même ligne de mot ont été effectuées. La Figure 3.35 montre une dispersion sur la même ligne de 13nm ce qui correspond à 10% de la largeur de grille visée.

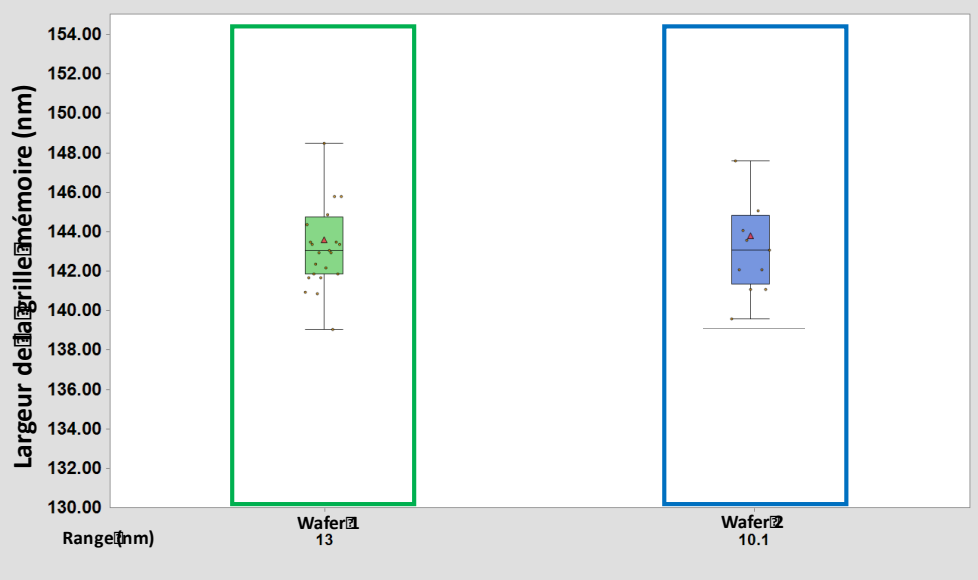


Figure 3.35 : Mesure de dispersion de la grille mémoire pour deux plaques du même lot

Cependant les mesures effectuées pour mettre à jour cette problématique sont assez complexes et ne peuvent pas être réalisées de manière automatique. Le challenge était de trouver un moyen d'automatiser cette mesure afin d'être capable de l'intégrer dans un environnement de production. L'état de l'art définit généralement la rugosité de bord de ligne par deux paramètres : le Line Edge Roughness (LER) et le Line Width Roughness (Figure 3.36) [78]–[80]. Cependant, l'impact électrique de la rugosité de bord sur le fonctionnement du transistor mémoire est évalué par le LWR. C'est pour cette raison que nous avons choisi de nous intéresser uniquement au LWR dans le cadre de cette étude.

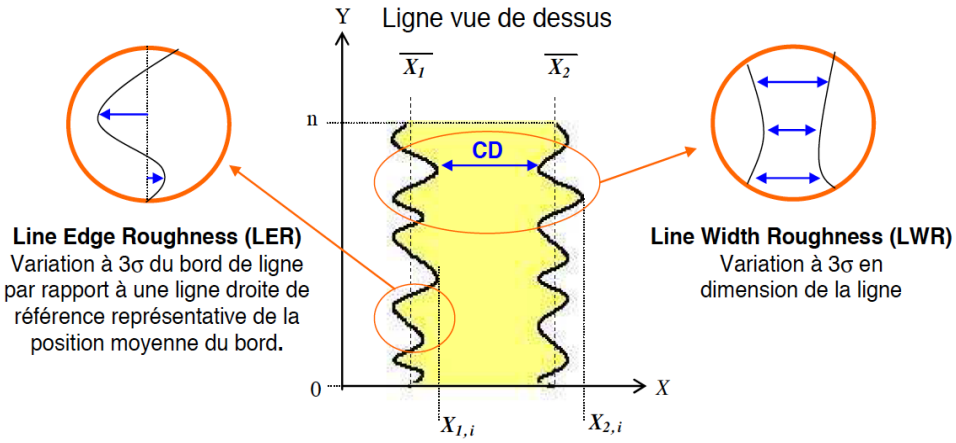


Figure 3.36 : Définition des deux indicateurs de rugosité de ligne : LWR et LER

Les mesures du LWR par l'intermédiaire des microscopes électroniques à balayage SEMCD sont basées sur une analyse de contraste d'image. Malgré une mise en œuvre aisée, cette méthode de mesure présente néanmoins certaines limitations :

- Le SEMCD fournit des images uniquement en deux dimensions, ce qui ne donne aucune information sur le profil de la structure comme le requiert l'industrie des semi-conducteurs. Les informations dimensionnelles à différentes hauteurs de la structure ne peuvent donc pas être obtenues.

- Le faisceau d'électrons endommage les profils de résine photosensible après photolithographie. Lors de l'acquisition de l'image, un phénomène de réduction en dimension de la résine est observé [81] modifiant la dimension réelle de la ligne et de la rugosité. Dans notre cas d'étude, il n'y aura pas d'impact sur la résine, car la mesure est effectuée après la gravure de la grille mémoire.

Malgré les limitations du SEMCD, les mesures effectuées ont montré l'existence d'une rugosité de l'ordre de 14nm en moyenne (Figure 3.37). Cette valeur est proche de ce qui a été trouvé avec les mesures manuelles (13nm). Cependant les limitations du SEMCD, utilisé sur le site de STMicroelectronics-Rousset, montrent les difficultés de cet équipement à effectuer une mesure précise de la rugosité de bord de ligne répondant aux exigences requises pour les transistors des futures générations. Par conséquent, le SEMCD a été utilisé seulement dans le cadre de ces travaux pour confirmer l'existence de la rugosité de bord.

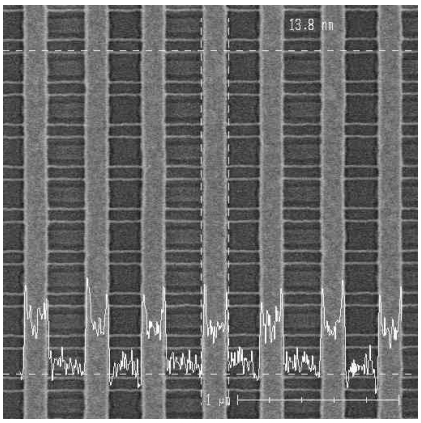


Figure 3.37 : Image SEMCD après mesure LWR

III. Conclusion

L'objectif de ce chapitre était de trouver une méthodologie d'un point de vue statistique qui permettra d'identifier, au premier ordre, les causes à l'origine des variations des paramètres électriques. L'état de l'art a montré que la majorité de ces méthodologies requièrent une grande quantité de données. Cependant, le fait que l'architecture eSTM soit en cours de développement ne permet pas d'utiliser une méthodologie standard pour analyser la variabilité du procédé de fabrication. Pour contourner ce problème, une étude approfondie a été effectuée des étapes critiques du procédé de fabrication pouvant impacter le bon fonctionnement de l'eSTM. Cette étude a été effectuée à l'aide de plusieurs campagnes de mesures en ligne pour évaluer la variabilité lot-à-lot, intra-wafer, intra-champ et intra-die.

A partir de cette étude nous avons pu quantifier les différentes variabilités du procédé de fabrication. Cela a permis de proposer une Correction Optique de Proximité pour corriger la variabilité intra-die de la largeur de la zone d'active et celle de la tranchée. Cette étude nous a aussi permis de mettre en évidence la rugosité des lignes de la grille mémoire ainsi que la variabilité temporelle de la largeur de la tranchée. Ces deux derniers paramètres ont un impact sur un autre paramètre critique de l'architecture eSTM : la distance entre la tranchée et la grille mémoire.

L'analyse théorique et les expériences réalisées dans ce chapitre ont montré l'impact de la dispersion de la distance entre la tranchée et la grille mémoire sur le bon fonctionnement de la cellule mémoire. Pour limiter la dispersion de cette distance, nous allons proposer une boucle de régulation entre la gravure de la tranchée et celle de la grille mémoire. Cependant, cette boucle de régulation ne peut être mise en place sans l'élimination de la rugosité des lignes de la grille mémoire. Pour cela, l'objectif du prochain chapitre sera d'étudier l'amélioration de la rugosité des lignes de la grille mémoire et la mise en place de la boucle de régulation.

Chapitre 4 : DEVELOPPEMENT ET MISE EN PLACE DE LA BOUCLE DE REGULATION

Sommaire

Ce chapitre est dédié à l'optimisation du procédé de fabrication de l'architecture eSTM afin d'améliorer les différentes variabilités mises en avant dans le chapitre 3. L'objectif principal est de mettre en place une boucle Run-to-Run (R2R) entre la largeur de la tranchée du transistor de sélection et la largeur de l'empilement mémoire. Cette boucle de régulation aura pour objectif de garantir une distance constante entre ces deux éléments. Dans un premier temps une nouvelle recette de gravure de la grille mémoire va être proposée pour améliorer la rugosité de la ligne observée dans le chapitre 3 (paragraphe II.6). Ensuite nous allons détailler le plan d'expériences utilisé pour la construction du modèle prédictif de la gravure du transistor mémoire. Puis l'implémentation de la boucle R2R et les résultats en ligne seront décrits, et pour finir l'étude de l'impact de la boucle de régulation sur la fiabilité de la cellule mémoire sera détaillée dans la dernière partie.

I. Amélioration de la rugosité de la ligne "la grille mémoire"

Les analyses effectuées dans le Chapitre 3 ont révélé une importante variation de la largeur de l'empilement mémoire le long de la ligne de mot (Figure 4.1). L'impact de cette variation sur les performances électriques s'est révélé important. L'objectif de cette partie est d'étudier l'opération de gravure de l'empilement mémoire. Cette étude nous permettra de trouver la source de cette rugosité, ainsi qu'une solution pour améliorer cette étape du procédé de fabrication.

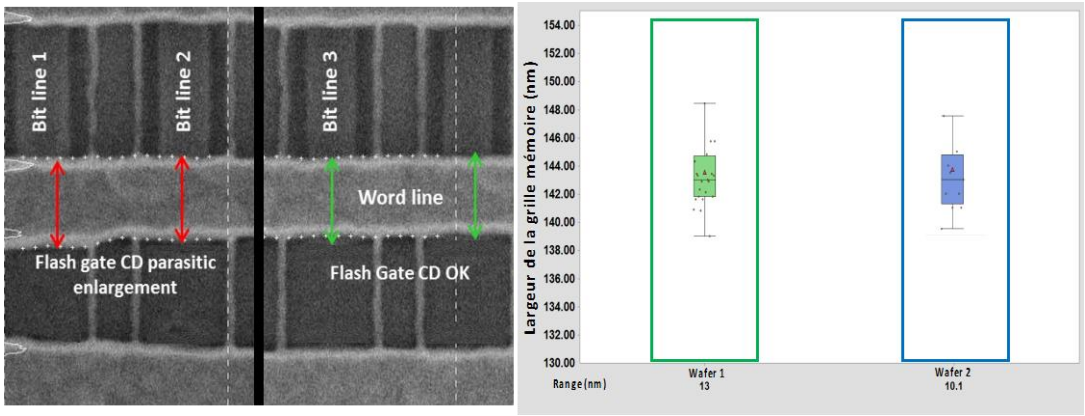


Figure 4.1 : Résultats d'analyse de variabilité

1. La gravure du transistor mémoire

La réalisation du transistor mémoire est une étape critique du procédé d'intégration des mémoires embarquées. Le dépôt d'une couche de BARC (*Bottom Anti Reflective Coating*) avant celle de la résine (Figure 4.2a) est impératif pour limiter les réflexions durant la photolithographie et d'assurer une bonne définition de l'empilement mémoire [82]. Après l'étape de photolithographie, l'opération de gravure du transistor mémoire de l'eSTM est auto-alignée et elle consiste à utiliser différents éléments chimiques spécifique pour chaque couche à graver. La Figure 4.2 détaille les étapes de cette gravure complexe faisant intervenir différents réactifs chimiques :

- De l'Argon (Ar) pour la gravure de la couche du BARC
- Un mélange de Bromure d'hydrogène (HBr) et du tétrafluorure de carbone (CF_4) pour la gravure des deux couches de polysilicium.
- Un mélange d'hélium et du CF_4 pour la gravure du tri-couche ONO.

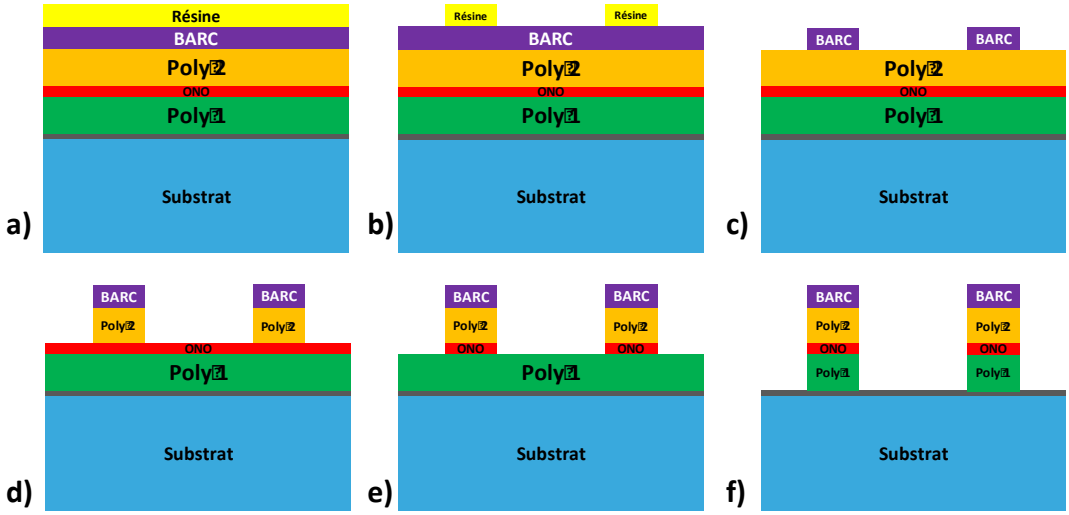


Figure 4.2 : Etapes de définition de la grille mémoire – a) Déposition du BARC et de la résine – b) Etape de photolithographie c) Gravure du BARC – d) Gravure du Poly1 – e) Gravure de l'ONO – f) Gravure du Poly2

L'utilisation du "*partitioning*" a été primordiale dans la compréhension de l'effet des différentes étapes de cette recette. Cette méthode consiste en une analyse physique après chaque couche gravée. Les images SEM de la Figure 4.3 montrent l'évolution de l'échantillon après chaque étape de l'opération de gravure. Il est important de noter l'apparition d'une déformation importante des grilles mémoires après la gravure de la couche de BARC (Figure 4.3a). Cependant ces déformations ont été partiellement absorbées pendant la gravure des autres couches. Après concertation avec l'équipe de l'atelier gravure, la rugosité de l'empilement peut être fortement améliorée en travaillant sur l'étape de gravure du BARC.

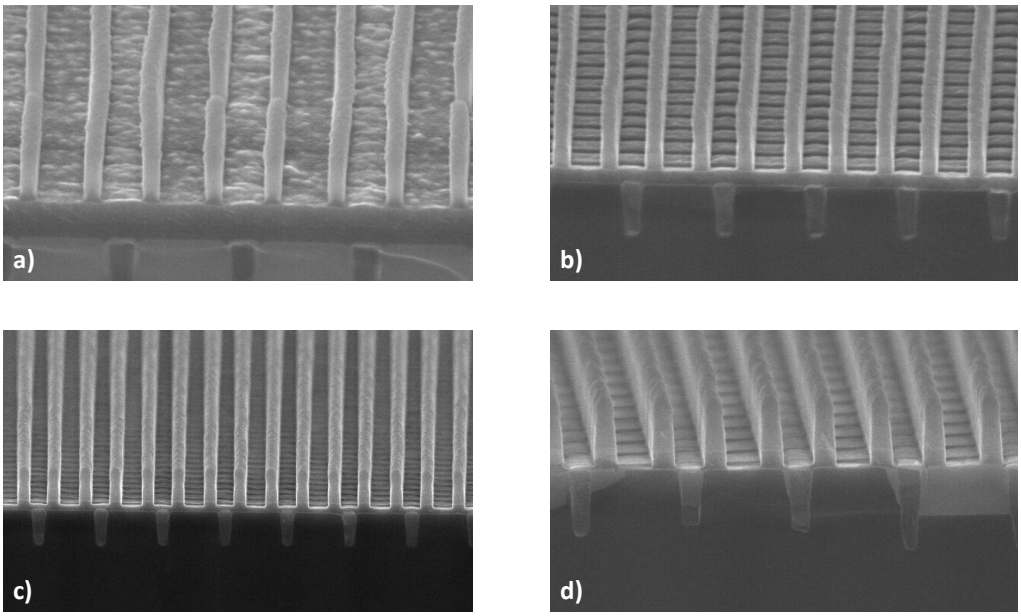


Figure 4.3 : Photos SEM du partitioning – a) Après gravure du BARC – b) Après gravure du Poly1 – c) Après gravure de l'ONO d) Après gravure du Poly2

Le partitioning a montré que le BARC était la source de la rugosité de la ligne de mot. Celle-ci pourrait-être due à l'utilisation d'une gravure physique pour le retrait de la couche de BARC [83] : la rugosité des lignes de résines serait projetée sur le BARC transférant ce problème

sur l'ensemble de l'empilement de grille mémoire. Pour limiter ce phénomène, l'élément chimique effectuant la gravure du BARC a été changé. Ce nouvel élément chimique qui est le tétrafluorure de carbone (CF_4) permettra d'atténuer la rugosité projetée sur les grilles mémoires (Figure 4.4).

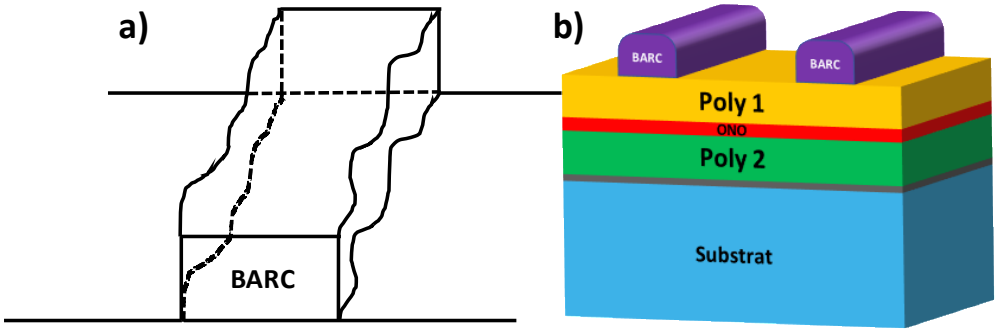


Figure 4.4 : Impact de la gravure sur les lignes de BARC – a) Gravure avec de l'Ar – b) Gravure avec le CF_4

2. L'impact de la nouvelle recette de gravure : tétrafluorure de carbone

L'objectif de cette partie est d'étudier l'impact de la nouvelle recette de gravure de la grille mémoire sur le bon fonctionnement de la mémoire eSTM. Un lot de 13 plaques a été dédié à cette étude limitant ainsi le coût des expériences. La première étape consiste à observer l'impact de cette recette sur la rugosité de la ligne de mot. Dans un deuxième temps sa répétabilité a été mise à l'épreuve. Enfin son effet a été évalué au niveau des paramètres électriques de la cellule.

a) Résultats en ligne

Dans le but d'évaluer la rugosité de la ligne de mot une campagne de mesure manuelle a été effectuée sur différents échantillons pour les comparer aux résultats montrés dans le chapitre 3 (paragraphe II.6). Comme l'illustre la Figure 4.5, la nouvelle recette permet une réduction de 70% de la dispersion de largeur d'une même ligne de mot par rapport à l'ancienne recette. Cette amélioration est due à l'utilisation du tétrafluorure de carbone pour l'élimination de la couche de BARC.

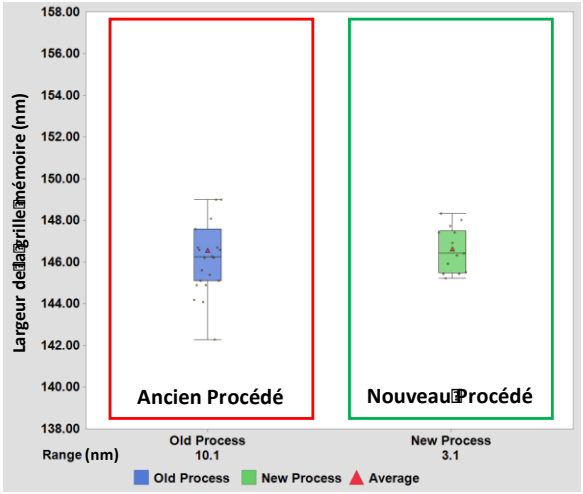


Figure 4.5 : Dispersion de la dimension critique de la grille mémoire sur une même ligne de mot pour le nouveau et l'ancien procédé de gravure

Il est important d'évaluer la répétabilité de ce nouveau procédé de gravure en cas de mise en production. La Figure 4.6 montre la mesure de la largeur de la grille mémoire sur 10 plaques du même lot. Il est important de noter qu'en plus de la stabilité de la valeur de la dimension critique, la variation intra-plaque est stable et acceptable.

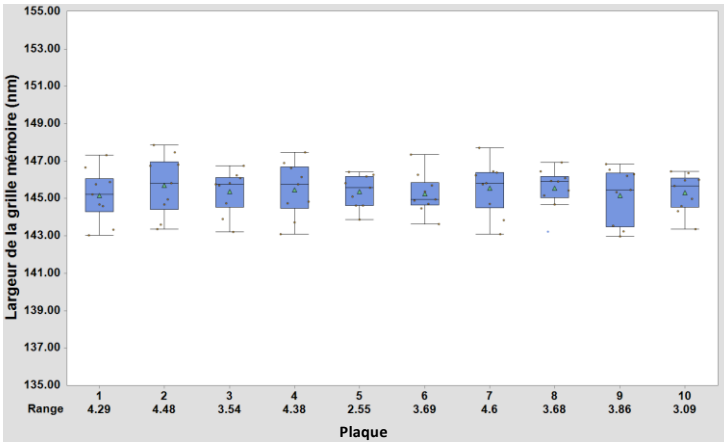


Figure 4.6 : Dispersion de la largeur de la grille mémoire après gravure pour évaluer la répétabilité de ce nouveau procédé de gravure

b) Résultats électriques

L'impact de ce nouveau procédé sur les paramètres physiques étant positif, il est temps d'évaluer son impact sur le bon fonctionnement électrique des dispositifs. Pour cela deux types de tests électriques ont été effectués sur le lot d'expériences. Le premier consiste à contrôler les différentes structures unitaires logées dans la ligne de découpe (*Scribe Line*). Le deuxième test a pour objectif d'évaluer l'impact du nouveau procédé sur le rendement final. Comme mentionné dans le Chapitre 3, la rugosité de la grille mémoire a un impact sur les performances électriques de la cellule eSTM, en particulier sur la tension de seuil effacée. La Figure 4.7 illustre l'amélioration de la tension de seuil après effacement du nouveau procédé de gravure comparée à l'ancien. Nous pouvons remarquer une diminution de 87% de la dispersion du V_{TE} par rapport à l'ancien procédé. La Figure 4.8 montre qu'il n'y a pas eu de retombée négative sur les autres indicateurs de performance clients comme le gain des cellules paire et impaire, celui du transistor de sélection ou la "*CAST*".

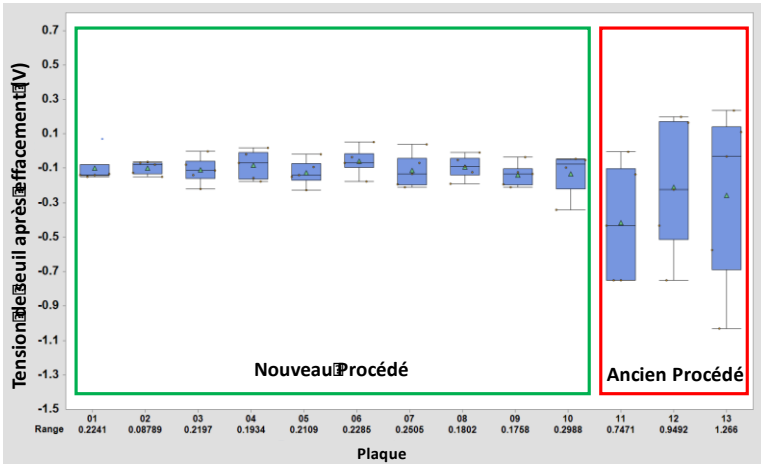


Figure 4.7 : Dipersion de la tension de seuil en fonction du procédé de gravure de la grille mémoire

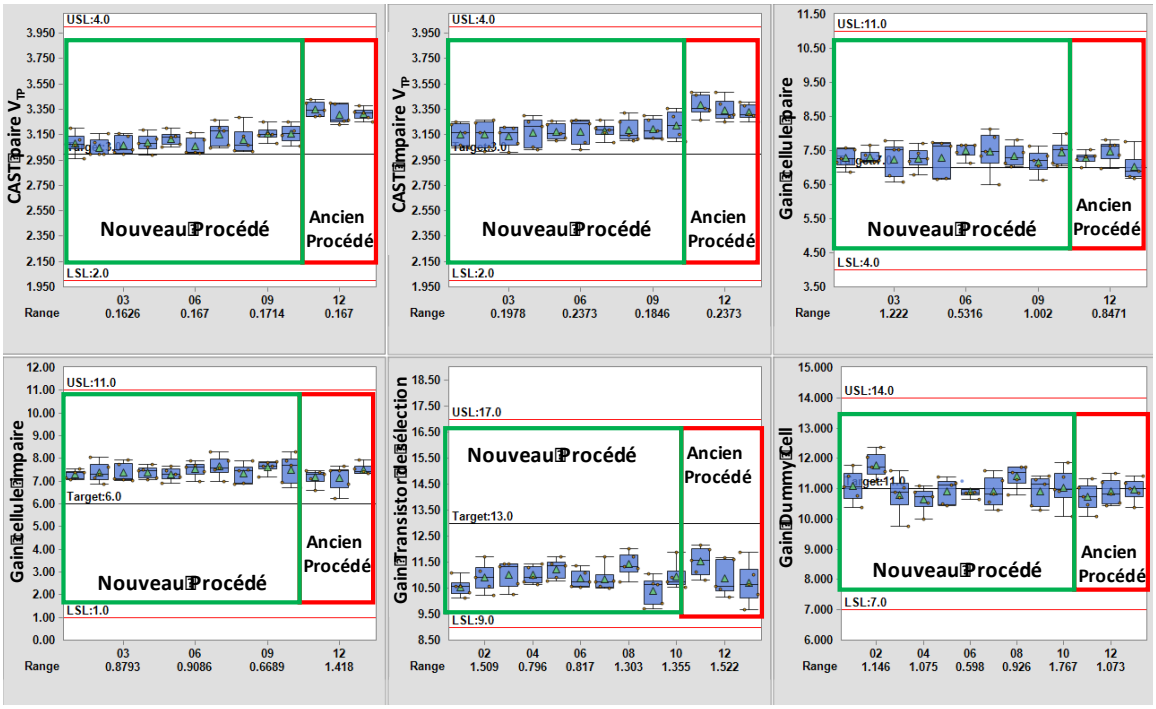


Figure 4.8: Impact du nouveau procédé de gravure de l'empilement mémoire sur les indicateurs de performances clients – Gain des cellules paire et impaire etc.

Enfin la Figure 4.9 montre une comparaison du rendement final entre le nouveau et l'ancien procédé de gravure de la grille mémoire. Le nouveau procédé de gravure permet une nette amélioration du rendement final, au-dessus de 90%, contrairement à l'ancien procédé qui avoisine 86%.

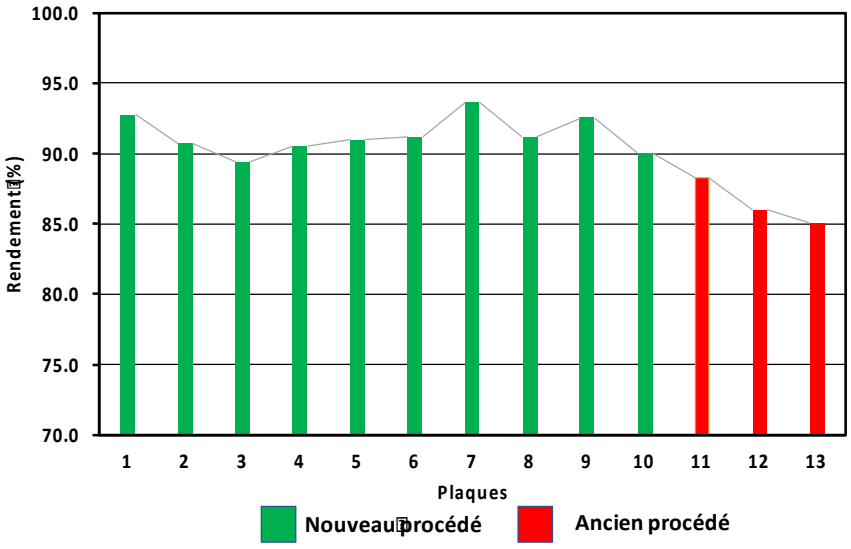


Figure 4.9 : Rendement du test électrique EWS en fonction du procédé de gravure de la grille mémoire

L'amélioration de la rugosité de la grille mémoire ainsi que les résultats électriques positifs nous permettent de remplacer le procédé de gravure. Ce changement de procédé va nous permettre de proposer une solution pour compenser la variabilité temporelle de la largeur de la tranchée (L_{TR}) identifiée dans le chapitre 3 (paragraphe II.5b).

II. Etude du développement et de la mise en place d'une boucle de régulation

Dans cette partie une régulation de compensation (*Feed-Forward Controller*) [84] sera développée entre la gravure de la tranchée du transistor de sélection et celle de l'empilement mémoire (Figure 4.10). L'objectif de cette régulation est de maintenir constante la distance entre ces deux éléments (L_D) et de compenser la variabilité de L_{TR}. Le contrôleur ajustera la largeur de la grille mémoire de chaque lot. En pratique, les lots ayant une largeur de tranchée surdimensionnée auront une grille mémoire plus faible et les lots ayant une largeur de tranchée sous-dimensionnée auront en revanche une grille mémoire plus grande.

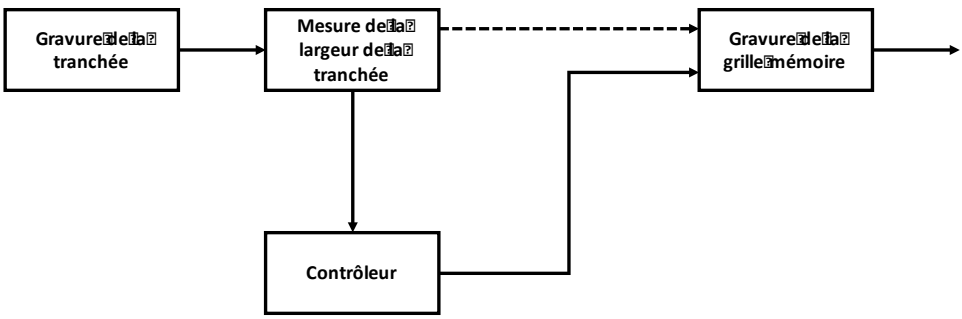


Figure 4.10 : Description de la boucle de régulation *Feed-Forward* réalisée entre l'étape de gravure de la tranchée du transistor de sélection et celle de l'empillement mémoire

Avant de décrire le fonctionnement du contrôleur de cette boucle de régulation et sa mise en place, il est nécessaire d'expliquer la méthodologie utilisée ainsi que les différents types de boucles de régulation existants dans l'industrie des semi-conducteurs.

1. Méthodologie

a) Types de boucles de régulation

Chaque étape du procédé de fabrication induit une variabilité au niveau des plaques. Cette variabilité peut s'exprimer de batch[8] à batch, de lot à lot, de plaque à plaque, de site à site [74]. Il existe des variabilités liées à l'équipement de mesure et enfin les variabilités liées à la nature du produit. La variabilité non expliquée constitue les résidus. On définit la variabilité totale comme étant la somme de ces dernières.

Afin de réduire ces variabilités, des boucles de régulations sont développées [85]. Le principe est de réajuster un ou plusieurs paramètres du procédé en temps réel en fonction des mesures physiques ou électriques des lots précédents. Ceci se fait en suivant un modèle développé à partir des expériences ou à partir d'un historique de données regroupant des mesures de métrologie. Il faut noter que la régulation est possible que pour les perturbations non aléatoires. Deux types de boucles de régulations peuvent être distinguées :

- Les boucles de régulation de type « Feed Back » sont déployées généralement sur un même équipement à partir des mesures des lots précédents (N-1). L'écart de la valeur mesurée de la plaque du lot (N-1) permet d'estimer la valeur du paramètre équipement pour le lot (N). La Figure 4.11 montre un exemple de boucle « Feed back » au niveau du procédé de Polissage Mécano-Chimique (CMP). La valeur de l'épaisseur d'oxyde enlevée est renvoyée au contrôleur pour calculer un nouveau temps de polissage qui sera utilisé pour le lot suivant.
- Les boucles de régulation de type « Feed Forward » sont utilisées pour rattraper une dérive observée sur l'étape de fabrication N en modifiant les paramètres de la recette de l'étape (N+1). Ce type de boucle est connu aussi sous le nom de "boucle de compensation". Par exemple, la Figure 4.11 illustre le cas d'une boucle entre le procédé de Dépôt Chimique en phase Vapeur et le procédé de Polissage Mécano-Chimique (CVD-CMP). La vitesse de polissage est modifiée de plaque à plaque suivant les valeurs des épaisseurs de dépôt de chaque plaque.

[8] Le batch est défini comme un ensemble de deux ou plusieurs lots. Par exemple lors d'une opération de diffusion, 6 lots sont disposés ensemble dans le four. On parle donc d'un batch de 6 lots.

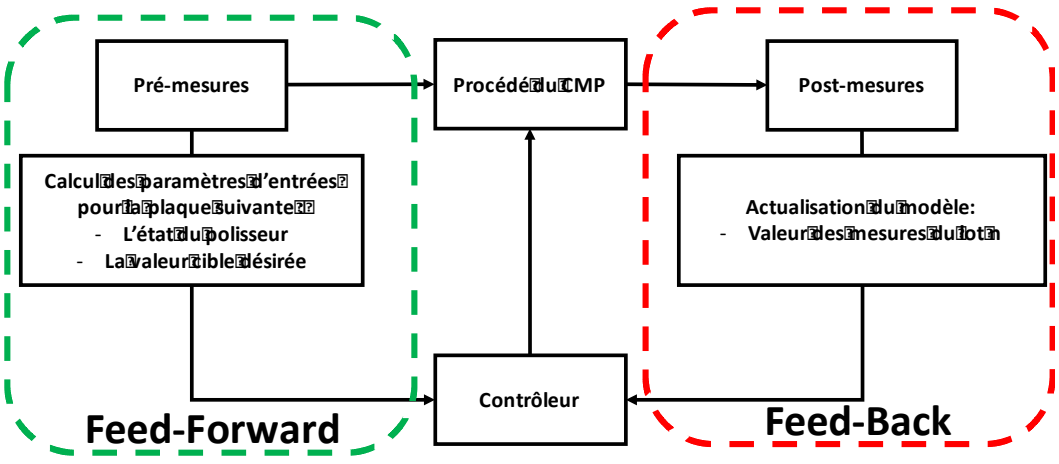


Figure 4.11 : Schéma de deux boucles de régulation Feed Back et Feed Forward – Source : STMicroelectronics.

b) Modèle de procédé prédictif

Pour chacune des boucles de régulation, un modèle de procédé est recherché. Suivant les paramètres de procédé ajustables et les mesures à contrôler, les modèles peuvent être de type [86] :

- SISO (Single Input Single Output) où il y a une seule variable d'entrée (un seul paramètre de procédé ajustable) et d'un seul paramètre de mesure.
- MIMO (Multiple Input Multiple Output) où il y a plusieurs variables d'entrée (plusieurs paramètres du procédé ajustables) et de plusieurs paramètres de mesure.
- MISO (Multiple Input Single Output) où il y a plusieurs variables d'entrée (paramètre de procédé) et un seul paramètre de mesure.
- SIMO (Single Input Multiple Output) où il y a une seule variable d'entrée (paramètre de procédé) et plusieurs paramètres de mesure.

Les modèles du procédé représentent le cœur du contrôleur. Il est recommandé d'avoir un modèle simple (linéaire) pour des raisons de simplicité mais aussi pour faciliter le transfert du modèle sur d'autres machines d'un même atelier et éventuellement vers d'autres FABs. Deux sortes de modèles sont distingués :

i) Les modèles issus des données de production

L'existence d'un nombre important de données sur un procédé de fabrication permet de faciliter l'obtention d'un modèle après un traitement approfondi. Cependant, l'étendu (*Range*) des différents paramètres physiques est bien contrôlé dans un environnement de production. Ce qui réduit l'étendue de leur variation, il est alors difficile de modéliser le comportement du procédé en dehors de ces plages habituelles des données de productions.

ii) Les modèles issus de plan d'expériences (DOE : Design Of Experiments)

Dans ce genre de modèle une fonction mathématique nous permet de relier la réponse aux facteurs. L'objectif est que le modèle prend la forme d'un polynôme de degré plus ou moins élevé :

$$Y = a_0 + \sum_{i=0}^{n} a_i X_i + \sum_{j=0}^{p} a_{ij} X_{ij} + \cdots + \sum_{i=0}^{n} a_{ii} X_i^2 + a_{ij\ldots z} X_i X_j \ldots X_z$$

Où :

- Y est la réponse ou la grandeur d'intérêt. Elle est mesurée au cours de l'expérimentation.
- X_i représente le niveau attribué au facteur i par l'expérimentateur pour réaliser un essai. Cette valeur est parfaitement connue.
- X_{ij} est l'interaction entre les différents facteurs.
- a_0, a_i, a_{ij}, a_{ii} sont les coefficients du modèle mathématique adopté. Ils ne sont pas connus et doivent être calculés à partir des résultats des expériences.

La modélisation de la réponse par un polynôme permet de calculer toutes les réponses possibles dans le domaine d'étude sans être obligé de faire les expériences [87], [88]. Chaque point d'expérience permet d'obtenir une valeur de la réponse. Le plan d'expérience fourni donc un système de *n* équations (*n* essais) à *p* inconnues (*p* coefficients) qui peut être représenté sous notation matricielle :

$$Y = aX + e$$

- Y est le vecteur des réponses
- X est la matrice du modèle
- a : le vecteur des coefficients
- e : est le vecteur d'écarts

Ce système est résolu en utilisant une méthode de régression basée sur le critère des moindres carrés. Il existe de nombreux logiciels qui exécutent ce calcul et donnent directement les valeurs des coefficients. Ce type de modèle est recommandé car il prend en considération les mécanismes physiques des procédés mais aussi les variations des paramètres en dehors des plages habituelles des données de production.

Il existe plusieurs plans d'expériences pour estimer un modèle, ci-dessous quelques exemples [89] :

- Les plans à deux niveaux (plans factoriels fractionnaires à deux niveaux, les plans de Koshal …).
- Les plans à plusieurs niveaux (plans complets à trois niveaux, Plans à niveau mixtes …).
- Les plans pour surface de réponse (les plans composites, plan D-Optimal …).

Dans notre cas d'étude, seul le plan pour surface de réponse sera utilisé car comparé aux autres plans d'expériences c'est celui qui permet de prédire la réponse avec une bonne qualité dans l'ensemble du domaine expérimental [90]. La deuxième raison est que ce modèle est le plus utilisé dans le domaine des semi-conducteurs [91].

➢ **Les plans composites**

Les plans pour surface de réponse permettent la construction des modèles mathématiques de second degré. Ils sont utilisés pour des variables continues. Pour deux facteurs, on a :

$$Y = a_0 + a_1X_1 + a_2X_2 + a_{12}X_1X_2 + a_{11}X_1^2 + a_{22}X_2^2 + e$$

Où X_1 et X_2 sont deux variables et X_1X_2 représente l'interaction entre ces deux variables.

La Figure 4.12 représente un plan composite à deux facteurs qui sera utilisé dans notre étude [92]. Le point N est le point central, il représente le point médian des deux facteurs. Durant l'expérience, ce point peut être répliqué une ou plusieurs fois pour évaluer la répétabilité du procédé. Les points A, B, C, et D sont les points expérimentaux comptant une modification simultanée des deux facteurs et les points E, F, G et H comptent une seule modification à la fois.

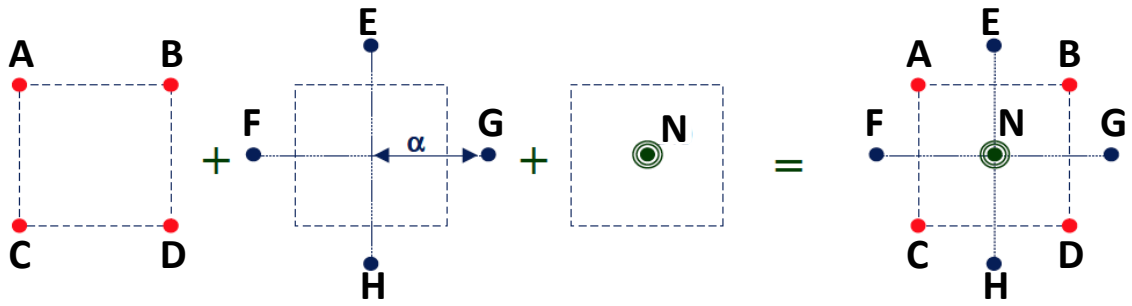


Figure 4.12 : Plan composite à deux facteurs [92]

➢ **Analyse statistique**

Après exécution des plans d'expériences, les résultats des mesures des expériences sont traités. Ainsi une analyse statistique est requise pour juger la qualité du DOE effectué. Dans notre cas, le logiciel Design Expert permettra de mettre en évidence l'erreur du modèle et les éventuelles corrélations que la construction du plan ne permet pas d'expliquer. Pour cela, il existe deux paramètres statistiques pour quantifier cette analyse :

- Le coefficient de détermination ajusté (R^2 ajusté : Adj-R^2) traduit le taux de variations expliqué par les effets retenus dans le modèle.
- La probabilité que le résultat est significatif, ce paramètre est connu comme la valeur de la probabilité (p-value). Généralement la p-value est comparée à un seuil préalablement défini, si la p-value est inférieure à ce seuil le résultat du test est considéré significatif. Sous le logiciel Design Expert, la valeur de seuil est de 0.05.

Puis un test physique et statistique sera effectué pour valider le modèle dans la plage des variations des facteurs étudiés.

➢ **Evaluation du gain**

La dernière étape consiste à évaluer le gain sur la réduction de la variabilité. Le déploiement des boucles R2R doit apporter sa contribution à l'amélioration du rendement. L'évaluation du gain est estimée au niveau :

- Des paramètres physiques : il s'agit généralement d'évaluer le gain de la réduction sur la variabilité des valeurs des paramètres physiques impactées par la boucle R2R. Un paramètre physique (largeur, profondeur, épaisseur…) doit être défini comme indicateur de suivi des boucles de régulations.
- Des paramètres électriques : l'objectif est d'observer l'impact de la boucle de régulation sur la variabilité des paramètres électriques. Pour cela, il est important de choisir le bon paramètre électrique (courant, tension, résistance…) comme indicateur de suivi.
- Du test final des puces : un apport positif de la boucle R2R sur le rendement final permettra de justifier leur déploiement.

Après avoir expliqué les différentes étapes à suivre pour la mise en place d'une boucle de régulation, nous allons maintenant exposer les résultats de la construction de notre boucle de compensation.

2. Boucle de compensation entre la tranchée du transistor de sélection et le transistor mémoire

a) Identification des paramètres critiques

Comme expliqué auparavant le contrôleur utilisé dans notre boucle de régulation (Figure 4.10) est constitué d'un modèle de prédiction qui permet d'ajuster la largeur de la grille mémoire. Pour construire ce modèle de prédiction, nous avons besoin d'identifier les paramètres critiques du procédé de gravure de la grille mémoire. Pour cela des expériences sur le procédé de gravure ont été effectuées en amont afin de déterminer les paramètres impactant la largeur de l'empilement mémoire. Après concertation avec l'équipe de l'atelier gravure il a été retenu que les paramètres les plus influents d'un point de vue connaissances métier sont :

- Le temps de sur gravure
- La pression
- La puissance RF (Coil RF Power)

Ces paramètres vont être utilisés pour changer les conditions de gravure de la couche du BARC, ce qui nous permettra de réguler la largeur de la grille mémoire selon nos besoins.

b) Modélisation du procédé de gravure de la grille mémoire

i) Plan d'expériences

Après avoir défini les paramètres critiques, l'étape suivante est la construction d'un plan d'expérience de type central composite à faces centrées. Le nombre de variables est de 3 (temps de sur gravure, pression et la puissance RF). Les valeurs de la pression varient de 4.5 à 6.5 mT, le temps de sur gravure (Over Etch OE) exprimé en pourcentage varie de 30 à 50%, et finalement la puissance RF varie de 150 à 250W. Le choix des plages de ces paramètres se justifie en grande partie par la limite de tolérance du fonctionnement de l'équipement. Le nombre d'expériences est de l'ordre de 20 dont 6 expériences pour tester la répétitivité du point central. Pour éviter les variabilités liées à l'équipement, nous avons forcé le passage sur une seule chambre de gravure. Le Tableau 4.1 résume les différentes conditions expérimentales :

Plaques	A : Puissance RF (W)	B : Temps (%)	C : Pression (mT)

1	250	40	5.5
2	150	30	6.5
3	250	50	4.5
4	200	40	5.5
5	200	40	5.5
6	150	30	4.5
7	200	40	5.5
8	150	50	4.5
9	200	40	5.5
10	150	40	5.5
11	250	30	4.5
12	200	40	5.5
13	200	30	5.5
14	200	40	6.5
15	200	40	4.5
16	150	50	6.5
17	250	50	6.5
18	250	30	6.5
19	200	50	5.5
20	200	40	5.5

Tableau 4-1 : Récapitulatif des conditions expérimentales du plan d'expériences

ii) Mesures

Pour cette expérience, une cartographie standard des points de mesures a été choisie car l'évaluation des effets centre/bords n'est pas utile. La représentation de la cartographie en forme spirale de 9 points est illustrée sur la Figure 4.13.

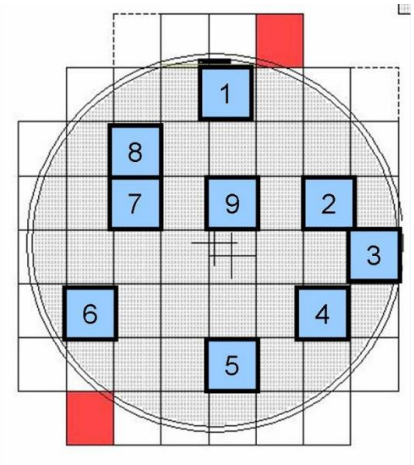


Figure 4.13 : Cartographie de mesures en forme de spirale de 9 points

iii) Résultats du plan d'expériences

L'analyse des résultats du DOE a été effectuée avec l'aide du logiciel Design Expert. Comme expliqué auparavant (paragraphe II.1), les résultats des indices statistiques sont déterminants pour juger la qualité du plan d'expériences effectué. Le Tableau 4.2 résume les

coefficients de détermination exprimant le taux de variations expliqué par les effets retenus, et le Tableau 4.3 donne les résultats des valeurs de p-value obtenues pour chaque paramètre :

Déviation Std	0.38	R^2	0.95
R^2 Ajusté	0.94	R^2 Prédit	0.87

Tableau 4-2 : Coefficients de détermination

Paramètre	P-value	Remarque
A : Puissance RF	<0.0001	Significatif
B : Temps	<0.0001	Significatif
C : Pression	0.014	Significatif
AB	0.0012	Significatif
BC	0.1082	Rejeté car P-value > 0.05
AC	0.1301	Rejeté car P-value > 0.05
A^2	0.0821	Rejeté car P-value > 0.05
B^2	<0.0001	Significatif
C^2	0.1828	Rejeté car P-value > 0.05

Tableau 4-3 : Valeurs de P-value obtenues sous Design Expert

À partir des résultats du Tableau 3.3, la largeur de la grille mémoire peut être exprimée par le modèle quadratique suivant :

$$Y = a_1A + a_2B + a_3C + a_4AB + a_5B^2 + \varepsilon$$

Avec : Y la largeur de la grille mémoire

a_1, a_2, a_3, a_4, a_5 sont les constantes associées aux interactions entre les trois paramètres

ε est une constante pour le centrage du modèle

Le modèle de la largeur de la grille mémoire est obtenu en fonction des paramètres du procédé de gravure ainsi qu'en prenant en compte leurs interactions. Maintenant, il est possible de comprendre l'impact de chaque paramètre critique sur la largeur de la grille mémoire. La réponse de surface de la largeur de la grille mémoire est donnée dans la Figure 4.14. Par exemple la largeur de la grille mémoire croît avec la décroissance de la puissance RF. Nous pouvons remarquer que le domaine de validité de notre plan d'expérience est restreint, ceci est dû aux choix des paramètres critiques du procédé de gravure. Ce domaine de validité est suffisant pour corriger la variabilité observée, mais il serait judicieux de l'améliorer pour une mise en production.

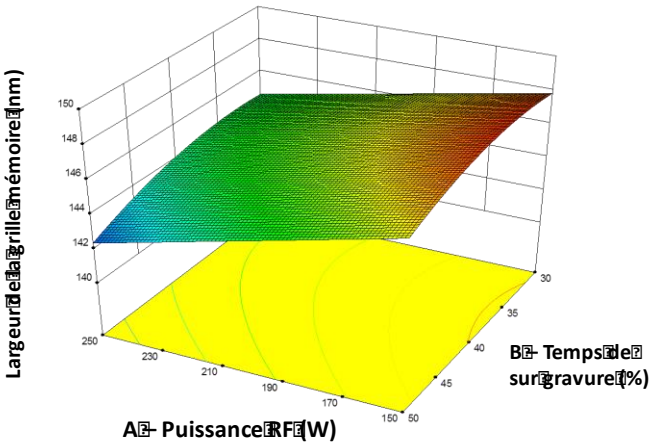


Figure 4.14 : Evolution de la largeur de la grille mémoire en fonction de la puissance RF (TCP) et du temps de sur gravure pour une valeur de pression de 5.5mT

iv) Simulation du modèle

Dans le but de valider le modèle établi, une série de simulations est indispensable pour tester le modèle. Le logiciel Design Expert permet la simulation de l'ensemble des combinaisons possibles en production. La Figure 4.15 illustre un exemple de cette interface.

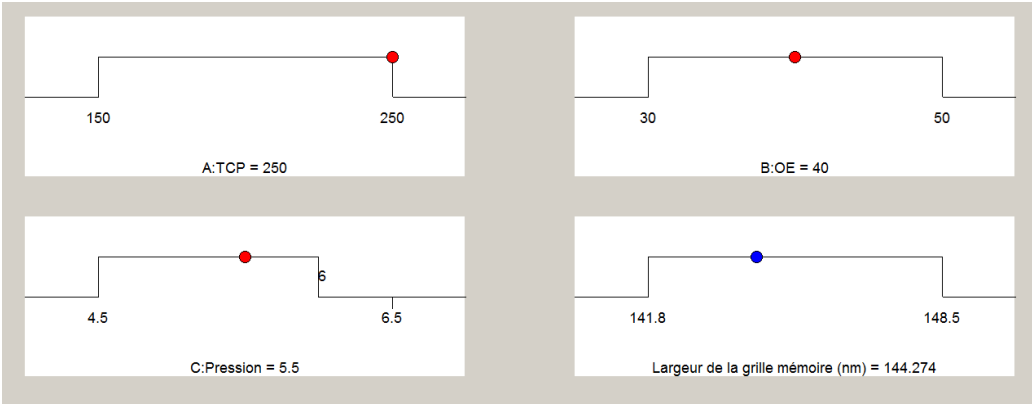


Figure 4.15 : Exemple d'une simulation à l'aide de Design Expert avec les conditions suivantes : TCP = 250W, Pression = 5.5mT, Temps OE = 40% pour une largeur d'empilement mémoire égale à 144.2nm

Comme nous pouvons le constater sur la Figure 4.15, nous pouvons simuler n'importe quelle largeur de grille mémoire à partir des conditions du procédé. Dans notre cas, il était impossible d'utiliser beaucoup d'expériences pour valider le modèle. C'est pour cette raison que le lot du plan d'expériences a embarqué 4 plaques avec des conditions différentes afin de valider le modèle. Cette expérience permettra de valider le modèle prédictif à moindre coût, le Tableau 4.4 résume les conditions des 4 plaques :

Plaques	Puissance RF (W)	Temps (%)	Pression (mT)
1	220	45	5
2	170	45	6
3	150	40	5
4	250	50	6

Tableau 4-4 : Récapitulatif des conditions expérimentales pour valider le modèle prédictif mise en place grâce au DOE

En simulant les mêmes conditions du Tableau 3.3, il est possible de comparer ces résultats avec ceux expérimentaux. Les résultats de cette comparaison sont illustrés sur la Figure 4.16. Le modèle présente une bonne prédictibilité face aux valeurs mesurées expérimentalement.

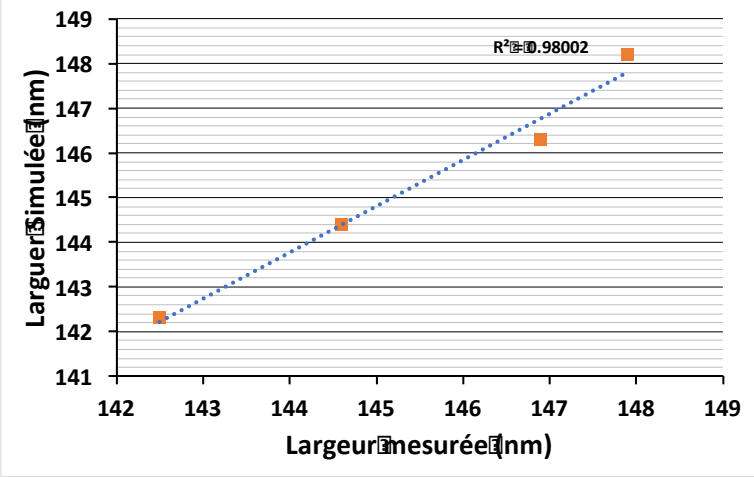


Figure 4.16 : Comparaison entre les données de simulation et les données expérimentales

Après avoir obtenu des résultats concluants avec la simulation, il est temps d'implémenter le modèle prédictif dans la boucle de régulation afin d'évaluer son impact sur le bon fonctionnement des dispositifs.

3. Implémentation de la boucle de régulation

Suite au transfert de la technologie eSTM sur le site de Crolles, il était impossible de tester notre boucle de régulation sur des lots de production. Pour résoudre ce problème nous avons proposé une expérience pour tester la boucle de régulation tout en minimisant les coûts. Cette expérience consiste à simuler la variabilité de la largeur de la tranchée (L_{TR}). Pour cela, nous allons prédéfinir la largeur de la tranchée en utilisant les limites de contrôles, deux lots pour la limite basse et haute et un lot standard. Ces limites de contrôles correspondent au pire cas de la variabilité tolérée lot à lot de L_{TR}. Nous avons utilisé 5 lots de 12 plaques et le Tableau 4.5 résume les conditions de passage au niveau de la photolithographie de la tranchée.

Lots	Conditions de passage
1	Limite haute de contrôle
2	Limite basse de contrôle
3	Largeur standard
4	Limite haute de contrôle
5	Limite basse de contrôle

Tableau 4-5 : Récapitulatif des conditions expérimentales pour simuler la variabilité de la largeur de la tranchée

La Figure 4.17 illustre les mesures de la largeur de la tranchée après l'étape de gravure pour chaque lot. Nous pouvons observer que la simulation de la variabilité lot à lot de L_{TR} est une réussite, maintenant il est temps d'utiliser la boucle de régulation pour compenser cette variabilité.

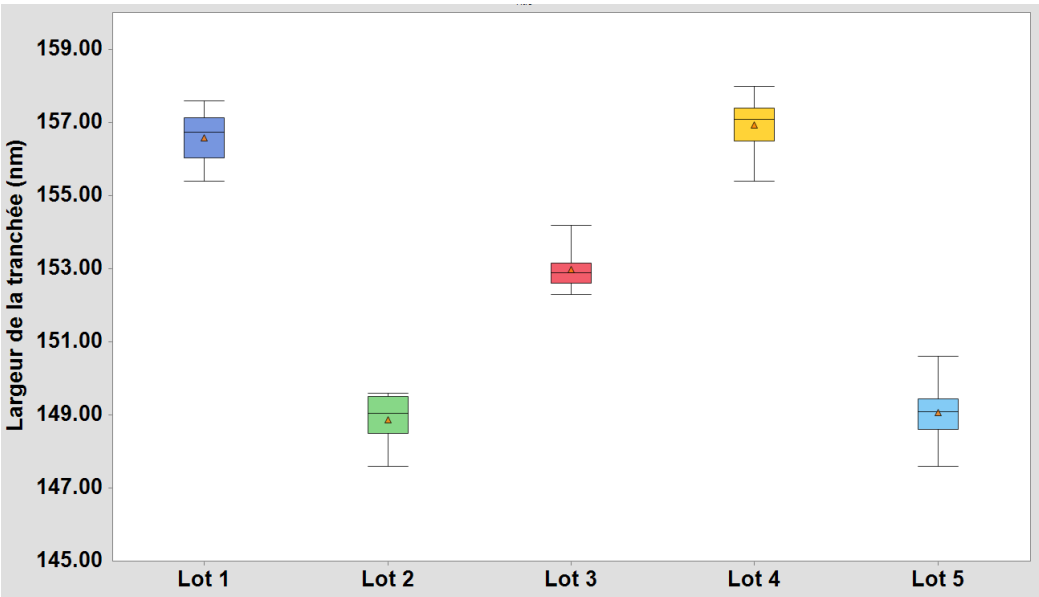


Figure 4.17 : Mesure de la largeur de la tranchée après gravure en fonction des lots d'expérience

Comme expliqué auparavant, l'objectif est de garder la distance entre la tranchée du transistor de sélection et l'empilement mémoire (L_D) constante. A chaque lot qui se présente au niveau de la gravure de la grille mémoire, le contrôleur doit fournir la largeur de la grille mémoire (L_G) qui permettra de compenser la variabilité de L_{TR}. Pour y parvenir, le module de prédiction ira chercher la valeur moyenne de L_{TR} par rapport au lot, ensuite il résoudra l'équation ci-dessous :

$$L_G = 2 * \left[K - \left(\frac{L_{TR}}{2} + L_{Dth}\right)\right]$$

Avec : K est le pitch, la distance entre le milieu de la tranchée et celui de l'empilement mémoire.

L_{Dth} : la valeur théorique de la distance entre la tranchée et la grille mémoire
L_{TR} : Largeur de la tranchée – L_G : Largeur de la grille mémoire

Pour étudier l'impact de la boucle de régulation sur la distance réelle entre la tranchée et la grille mémoire, nous avons comparé deux groupes de lots. Le premier a été traité avec la boucle de régulation et le deuxième sans cette dernière. Le Tableau 4.6 montre les conditions calculées avec le contrôleur pour corriger la largeur de la grille mémoire.

Lots expérimentaux	Puissance RF (W)	Temps (%)	Pression (mT)
Lot 1	250	50	6
Lot 2	150	40	6.5
Lot 3	200	40	5
Lot 4	250	50	4.5
Lot 5	150	40	6.5

Tableau 4-6 : Récapitulatif des conditions de gravure de l'empilement de grille du transistor mémoire

Suite à la réalisation des différentes expériences définies, les lots sont gravés et mesurés en ligne avec le SEMCD. La Figure 4.18 montre l'évolution de la distance entre la tranchée et la grille mémoire avec ou sans la boucle de régulation. Nous constatons une nette réduction de la

variabilité et un recentrage sur la valeur cible de L_D pour les lots traités avec la boucle de régulation.

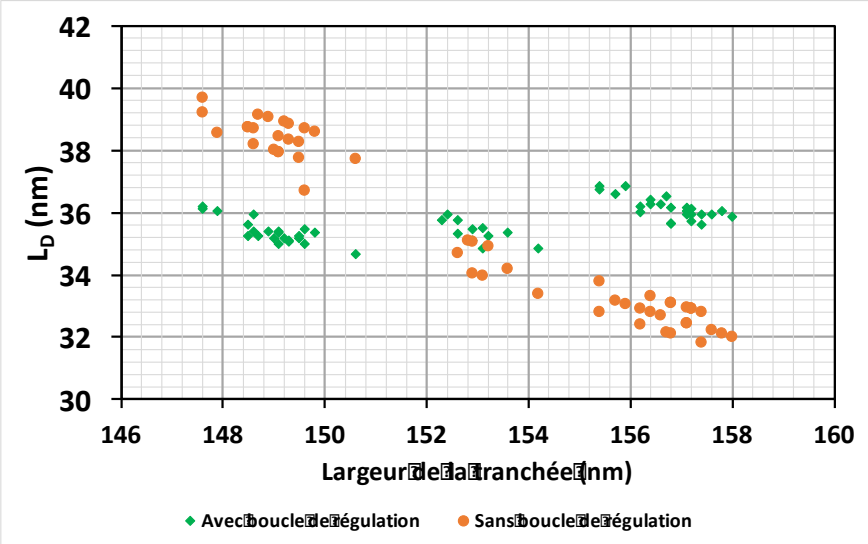


Figure 4.18 : Variation du paramètre L_D avec ou sans boucle de régulation en fonction de la largeur de la tranchée du transistor de sélection

Après ces résultats positifs au niveau des paramètres physiques, une analyse des performances électriques est primordiale pour évaluer l'impact de la boucle de régulation sur le bon fonctionnement de l'eSTM. Dans un premier temps, nous verrons l'impact sur les tensions de seuils V_{TE} et V_{TP} extraites du test paramétrique. Puis une analyse d'endurance sera présentée dans la dernière partie de ce chapitre afin d'évaluer les performances des dispositifs.

Comme mentionné dans le chapitre 3, une variation de la distance L_D modifie les tensions de seuils effacées et programmées. La Figure 4.19 illustre l'évolution du V_{TE} des lots avec la boucle de régulation comparée à un lot sans compensation. Nous constatons une réduction de la variabilité pour les lots avec la boucle de régulation et un recentrage du produit sur la valeur cible. En ce qui concerne la tension de seuil de programmation V_{TP}, on remarque sur la Figure 4.20 une nette amélioration de la dispersion intra-lot équivalente à celle du V_{TE}.

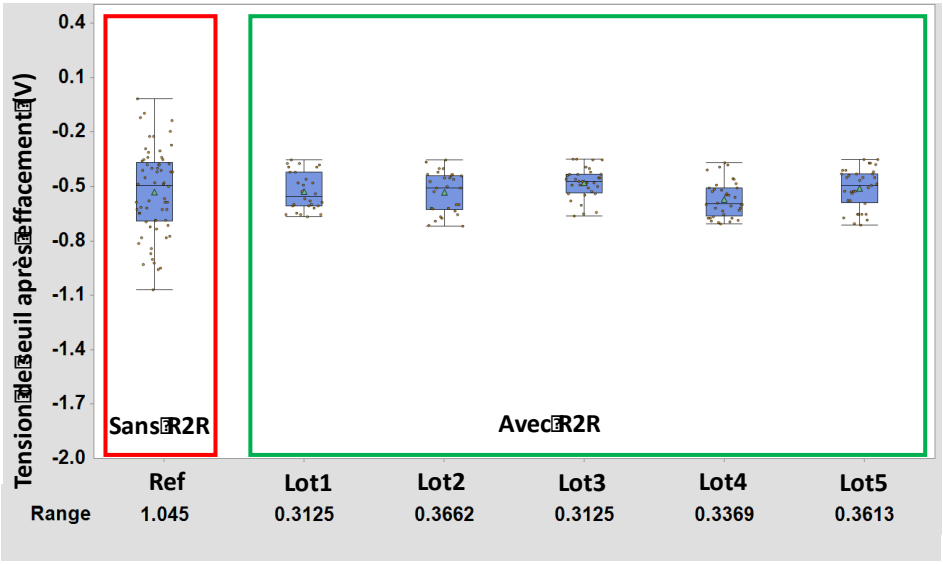


Figure 4.19 : Dispersion du V_{TE} pour les lots avec ou sans boucle de régulation

Figure 4.20 : Dispersion de la tension V_{TP} pour les lots avec ou sans boucle de régulation

L'impact du fonctionnement de la boucle de régulation sur les tensions de seuils effacées et programmées est très positif. La dispersion observée de ces tensions est de l'ordre de 400mV au maximum ce qui montre un gain de l'ordre de 60% avec la boucle de régulation. Conformément à l'analyse théorique faite dans le chapitre 3 (paragraphe II.5), le maintien du paramètre L_D constant permet de garantir un meilleur fonctionnement de la cellule eSTM. Au-delà de l'amélioration de la dispersion des tensions de seuils, nous avons analysé l'évolution des autres paramètres électriques de la cellule mémoire. La Figure 4.21 montre qu'il n'y a pas eu de retombée négative sur les autres indicateurs de performance "clients" comme le gain des cellules paire et impaire, celui du transistor de sélection ou la "dummy cell". L'analyse de l'impact de la boucle de régulation sur la fiabilité de la cellule eSTM sera développée à la fin de ce chapitre.

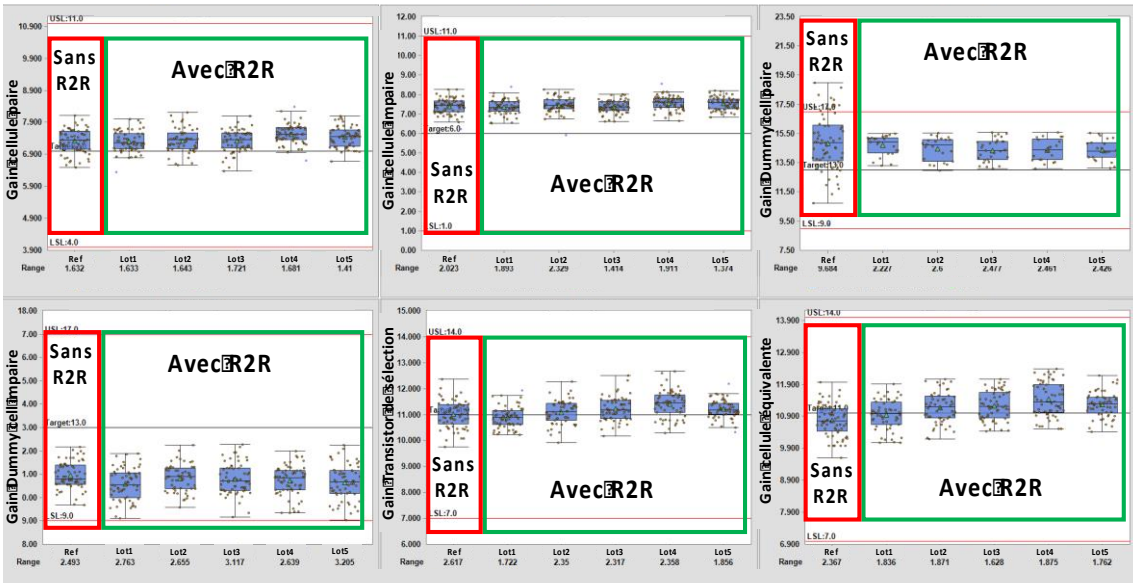


Figure 4.21 : Impact de la boucle de régulation sur les indicateurs de performances clients – Gain des cellules paire et impaire, celui du transistor de sélection et la dummy cell

Malgré le gain en variations de type lot à lot des tensions de seuils V_{TE} et V_{TP}, la mise en production de cette boucle de régulation à ce stade n'est pas possible. Car notre boucle de régulation ne répond pas aux exigences de l'équipe d'automation de STMicroelectronics-Rousset. Dans la prochaine partie, plusieurs actions vont être proposées pour améliorer la robustesse de la boucle de régulation.

4. Amélioration de la boucle de régulation

La mise en production du modèle prédictif consiste à implanter l'algorithme (équations, limites, filtre) dans une application dédiée aux régulations Run-to-Run. Cette application, Process Works de la société Rudolph Technology, est configurée pour communiquer avec les équipements de gravure et ceux de la métrologie. Le diagramme dans la Figure 4.22 illustre les différentes voies de communications entre l'application, la base de données de production (APC[9]), et le parc d'équipements. Pour chaque lot chargé sur un équipement de procédé, auquel ProcessWorks est associé, l'application va chercher différents éléments de son contexte de fabrication (route, technologie, opération, etc) à partir de la base APC. Seuls les lots concernés par les stratégies exigeant une régulation sont considérés et donnent lieu à un échange d'informations entre l'équipement de métrologie, le procédé d'un côté et l'application ProcessWorks de l'autre côté. Cette dernière stocke un certain nombre de variables dont les ajustements de la recette dans une base de données. A noter que la régulation est complétement automatisée et transparente aux yeux des opérateurs. Les utilisateurs ont toutefois accès à l'historique de la base de données via une interface dédiée.

[9] La base de données APC est mise à jour toutes les huit heures.

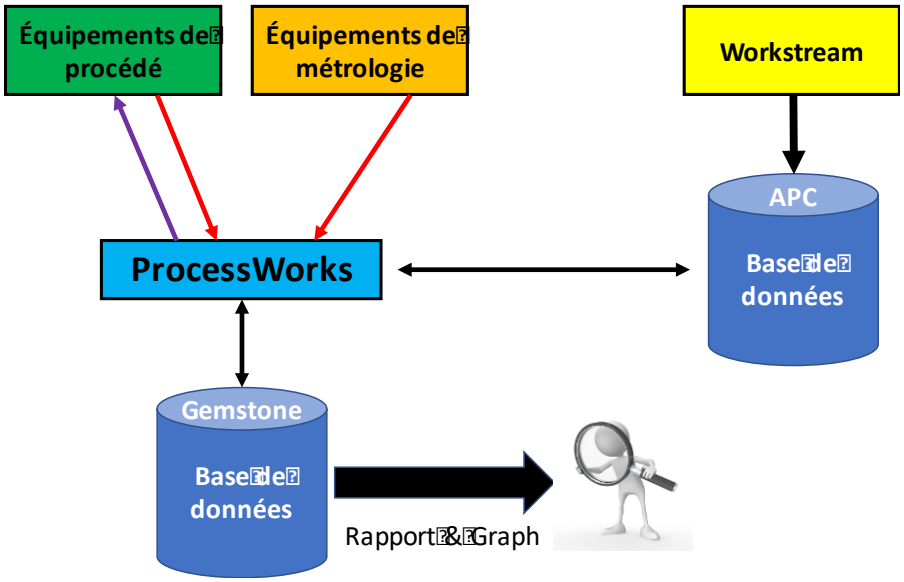


Figure 4.22 : Vue schématique de l'environnent de l'application ProcessWorks. Source : ST-Rousset

En se basant sur l'état de l'art [84], [85] et l'expérience des ingénieurs de ST-Rousset, deux éléments majeurs peuvent nuire à la bonne performance du contrôleur. Tout d'abord, le bruit de la mesure de la variable d'entrée. Un contrôleur qui s'appuie sur une mesure de métrologie trop bruitée, au-delà d'un certain seuil, pourrait en effet amplifier la variance de la variable de sortie. Le risque est de sur ajuster. En réponse à ce problème, le développement d'un filtre qui garantit la fiabilité de la mesure est primordial.

En seconde position figure l'interaction de la variable de commande avec des paramètres caractéristiques du produit, autres que le paramètre de sortie régulé. Prenons l'exemple de la régulation Feed-Forward entre les étapes de photolithographie et la gravure de la grille. Son principe de fonctionnement est simple, il s'agit d'ajuster les paramètres de gravure selon la déviation de la longueur de la résine mesurée en photolithographie. Cette méthode, bien que bénéfique pour la variabilité de la largeur de la grille, pourrait dégrader le profil de la grille. Afin de remédier à cela, il est possible de créer un groupe de recettes prédéfinies et qualifiées qui permettra de compenser la variabilité de largeur de résine.

En se basant sur le premier critère, notre boucle de régulation a besoin d'un filtre qui permet de vérifier la robustesse de la mesure de la largeur de la tranchée (Figure 4.23). Cependant, la métrologie utilisée (SEMCD) ne fournit que la valeur de la largeur de tranchée comme paramètre de sortie mais aucune indication sur la qualité de la mesure. L'utilisation de la scattérométrie comme méthode de mesure permettrait d'obtenir plusieurs informations liées à la mesure et surtout l'indicateur *Goodness Of Fit* (GOF). La valeur de cet indicateur est comprise entre 0 et 1. Elle est calculée pour chaque mesure individuelle en scattérométrie. Elle rend compte de la qualité de l'ajustement du modèle à la réponse scattérométrique mesurée par l'éllipsomètre. En général, une valeur de GOF inférieur à 0.95 signifie que la mesure n'est pas à prendre en compte. En utilisant l'indicateur GOF, il est possible d'introduire un filtre améliorant la robustesse de notre boucle de régulation. Aussi il serait essentiel de remplacer le SEMCD par la scattérométrie comme équipement de métrologie pour la mesure de la largeur de la tranchée.

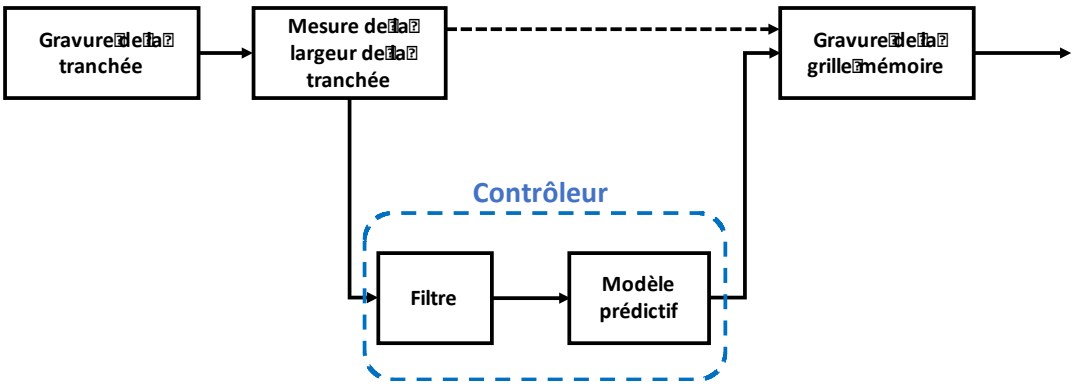


Figure 4.23 : Boucle de régulation avec filtre

En plus de l'ajout du filtre qui permettra de garantir la fiabilité de la mesure L_{TR} il est possible, de transformer le contrôleur en un régulateur adaptatif. Par définition, un régulateur adaptatif est un régulateur muni de coefficients ajustables permettant de maintenir un certain niveau de performances. Ce type de commande peut être appliqué à notre boucle de régulation (Figure 4.24), ce qui permettrait de corriger les dérives du procédé dans le temps.

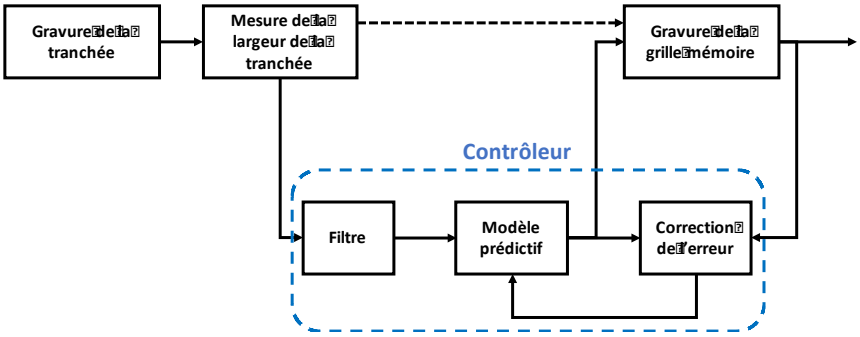


Figure 4.24 : Descriptif d'une boucle de régulation auto adaptative

Dans le but de rendre cette boucle de régulation auto adaptative [93], il est primordial d'étudier le risque d'instabilité qui peut être dû à l'inflation de la variance suite à des estimations en ligne biaisées. Pour cela, un nouveau modèle prédictif peut être envisagé avec une possibilité d'ajuster le centrage du modèle en se basant la mesure de la largeur de l'empilement mémoire.

III. L'impact du procédé de fabrication sur les performances de l'eSTM

L'objectif principal de cette partie est d'étudier l'impact des étapes critiques de la fabrication de la mémoire eSTM sur les performances électriques de cette dernière. Dans un premier temps, l'impact de la variation de la tranchée sur l'endurance de la cellule eSTM sera étudié. Ensuite l'effet de la rugosité de la grille mémoire, étudié dans le Chapitre 3, sur la variabilité intra plaque de l'endurance de la cellule eSTM sera présenté. Pour finir, l'efficacité des optimisations de procédé de fabrications proposées dans ce chapitre, sera évaluée pour améliorer la fiabilité de la cellule eSTM.

1. L'impact de la variation de la largeur de la tranchée

Comme expliqué auparavant la distance entre le transistor mémoire et le transistor vertical a un impact direct sur l'efficacité de programmation ainsi que sur la tension de seuil de l'état

effacé. Pour évaluer l'impact de cette variabilité sur l'endurance de la cellule eSTM, nous avons repris trois plaques du lot utilisé dans paragraphe II.5b du Chapitre 3. Le Tableau 4.7 résume les valeurs de la largeur du transistor de sélection (L_{TR}) pour chaque plaque.

Plaques	Largeur de tranchée (nm)
1	162.8
2	175.3
3	185.5

Tableau 4-7 : Valeurs de la largeur de la tranchée du transistors de sélection (nm)

Pour cette expérience un test d'endurance de 500k cycles à température ambiante a été effectué. La Figure 4.25 représente la tension de seuil effacée et programmée en fonction du nombre de cycles de la première plaque (L_{TR} = 162.8nm), ainsi que des courbes $I_D(V_{GC})$ après chaque cycle de programmation et effacement. Nous pouvons remarquer une tension de seuil V_{TP} initiale de l'ordre de 5V, ceci s'explique par la proximité des cellules mémoire de la zone de génération des porteurs chauds. La Figure 4.25a montre aussi une dégradation très importante de la tension de seuil V_{TE} et V_{TP} durant le test d'endurance qui impacte l'évolution de la fenêtre de programmation. Nous remarquons sur la Figure 4.25b un décalage de la caractéristique $I_D(V_{GC})$ de l'état effacé et une dégradation de la pente sous le seuil après un grand nombre de cycle.

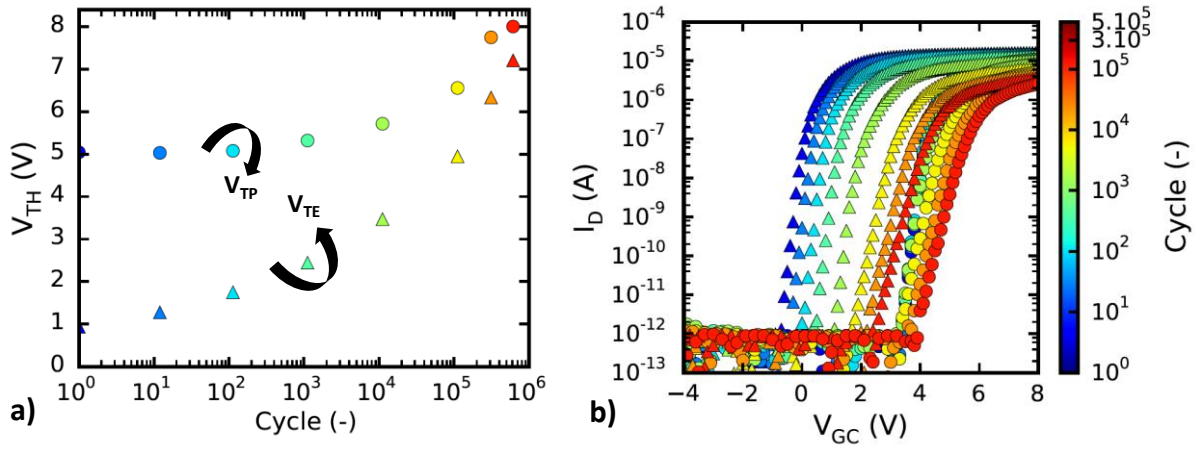


Figure 4.25 : a) Evolution des tensions V_{TP} et V_{TE} durant un test d'endurance de 500k cycles – b) Evolution des caractéristiques $I_D(V_{GC})$ en fonction du nombre de cycles pour la plaque 1.

La deuxième plaque (L_{TR} = 175.3nm) est considérée comme une plaque de référence du procédé de fabrication de la cellule eSTM. La Figure 4.26a illustre l'évolution des tensions de seuil programmées et effacées durant le test d'endurance. Nous remarquons une tension de seuil V_{TP} initial de l'ordre de 4.5V et une dégradation moins importante de la tension V_{TE} comparée à la première plaque. Ceci s'explique par la présence d'un implant flottant plus important qui retarde l'apparition des charges négatives dans l'oxyde tunnel.

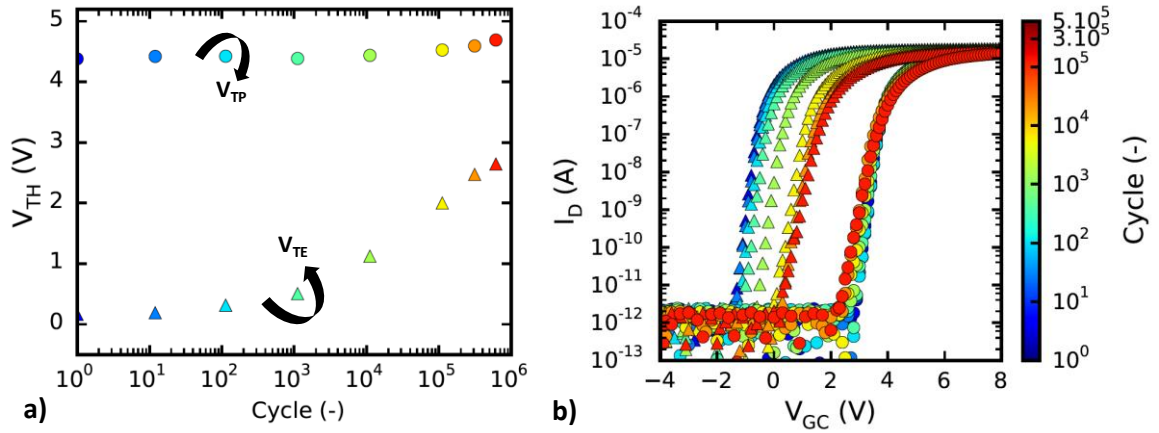


Figure 4.26 : a) Evolution des tensions V_{TP} et V_{TE} durant un test d'endurance de 500k cycles – b) Evolution des caractéristiques $I_D(V_{GC})$ en fonction du nombre de cycles pour la plaque 2

La caractérisation de la troisième plaque (Figure 4.27) montre une tension de seuil V_{TP} de l'ordre de 3.8V, ceci s'explique par une perte d'efficacité de programmation due à un implant très important entre la tranchée et le point mémoire. Cet implant flottant diminue l'énergie des porteurs chauds lors de la programmation réduisant l'injection. Néanmoins la dégradation de la tension de seuil effacée est équivalente à celle observée pour la deuxième plaque.

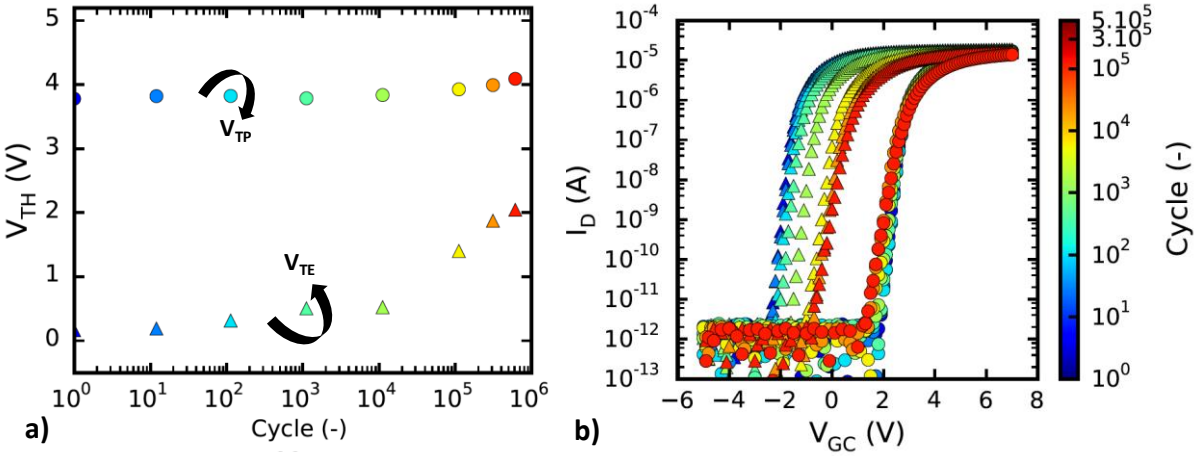


Figure 4.27 : a) Evolution des tensions V_{TP} et V_{TE} durant un test d'endurance de 500k cycles – b) Evolution des caractéristiques $I_D(V_{GC})$ en fonction du nombre de cycles pour la plaque 3

En comparant l'évolution de la fenêtre de programmation en fonction du nombre de cycles des trois plaques (Figure 4.28) nous remarquons, dans un premier temps une diminution de la fenêtre de programmation pour la première plaque à partir du 100ème cycle. Cette dégradation prématurée est due principalement à l'absence d'un implant flottant limitant l'énergie des porteurs chauds. De l'autre côté la dégradation de la fenêtre de programmation des deux autres plaques commence à partir de 1000 cycles car leur implant flottant retarde l'apparition des charges négatives dans l'oxyde tunnel et des défauts dans l'interphase oxyde tunnel / canal. La Figure 4.28 montre ainsi qu'une largeur de tranchée mal contrôlée peut induire une perte de plus de 70% de la fenêtre de programmation initiale à la fin du test d'endurance.

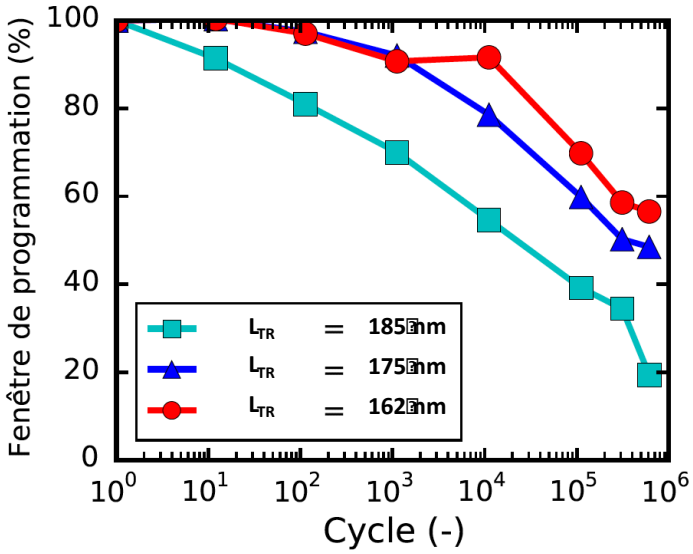


Figure 4.28 : Evolution de la fenêtre de programmation durant le test d'endurance pour les différentes largeurs de tranchée.

2. L'impact de la rugosité des grilles mémoires

Comme expliqué dans le chapitre 3, l'analyse de variabilité effectuée sur le procédé de fabrication de la cellule eSTM a montré une importante rugosité des grilles mémoires. Pour mieux comprendre l'impact de cette rugosité sur le bon fonctionnement de l'eSTM, nous avons effectué des tests d'endurance sur plusieurs dispositifs d'une même plaque. L'objectif est d'évaluer l'impact sur la variabilité intra-plaque.

La Figure 4.29 illustre l'évolution des tensions de seuil V_{TE} et V_{TP} en fonction du nombre de cycles. Les couleurs représentent la distance par rapport au centre de la plaque. Le test d'endurance montre une variabilité intra-plaque de la tension de seuil V_{TP} de l'ordre de 500mV. Cependant, la variabilité de la tension de seuil V_{TE} avoisine les 2V sans corrélation avec la distance centre bord des dispositifs. Ces variations confirment les corrélations effectuées dans le Chapitre 3 entre le test paramétrique et les mesures en ligne. La variation intra-plaque de la fenêtre de programmation engendre une perte en fiabilité très importante.

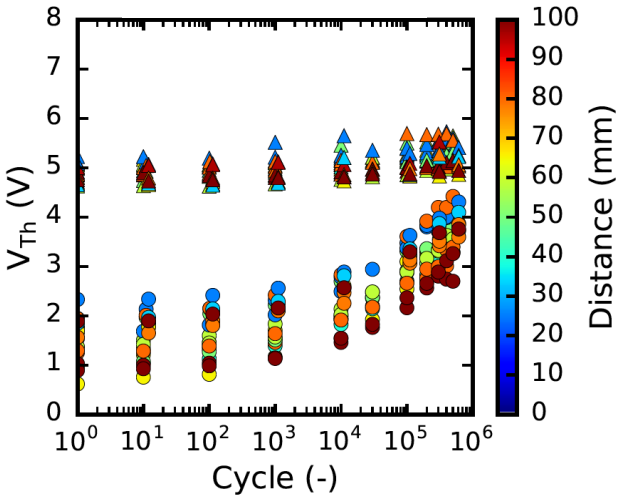


Figure 4.29 : Dispersion intra-plaque des tensions V_{TE} et V_{TP} durant le test d'endurance

Dans le paragraphe I de ce chapitre, un nouveau procédé de gravure du transistor mémoire a été proposé pour améliorer la rugosité des grilles mémoires. Ce nouveau procédé a permis une réduction de cette rugosité de l'ordre de 70%. Un test d'endurance a été effectué sur une plaque utilisant le nouveau procédé de fabrication. La Figure 4.30 montre l'évolution des tensions de seuil V_{TE} et V_{TP} en fonction du nombre de cycles. Nous remarquons une variation intra-plaque de l'ordre de 200mV des tensions de seuil V_{TP} et V_{TE} avec une dégradation de la fenêtre de programmation acceptable (10 dispositifs ont été testés).

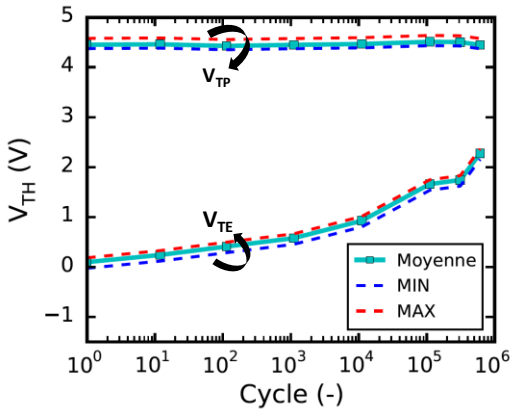


Figure 4.30 : Dispersion intra-plaque des tensions V_{TE} et V_{TP} durant le test d'endurance après amélioration du procédé de gravure

L'analyse des mesures effectuées dans cette partie montre que le nouveau procédé de fabrication améliore la variabilité intra-plaque de la fenêtre de programmation de la cellule eSTM. Avec ce résultat, il est possible maintenant de vérifier l'impact de la boucle de régulation entre la gravure de la tranchée et celle du transistor mémoire sur l'endurance de la cellule mémoire.

3. L'impact de la boucle de régulation

Dans le but de tester la boucle de régulation tout en minimisant les coûts, nous avons simulé la variabilité de la largeur de la tranchée. Cette expérience nous a permis d'évaluer l'impact de la boucle de régulation sur les performances électriques de la cellule mémoire. L'analyse du test paramétrique a montré une amélioration notable des différents paramètres électriques de la

cellule eSTM, l'objectif de cette partie est d'évaluer le gain par rapport à la fiabilité de l'eSTM. Dans un premier temps une étude de la variabilité lot à lot de la fenêtre de programmation sera présentée pour évaluer le gain de la boucle de régulation. Ensuite, nous verrons si la boucle de régulation a un impact négatif sur la variation intra-wafer de la fenêtre de programmation.

a) La variation lot à lot de la fenêtre de programmation

Les conditions de passage des lots utilisés pour le test de la boucle de régulation sont illustrées dans le Tableau 4.8. Les résultats en ligne (Figure 4.18) ont montré que la boucle de régulation a compensé la variabilité de la largeur de la tranchée avec l'ajustement de la largeur de l'empilement mémoire. Cette partie permettra d'évaluer la variabilité lot à lot des tensions de seuil programmées et effacées pour les cellules paire et impaire en fonction du nombre de cycle.

Lots	Largeur de la tranchée	Largeur de grille mémoire
1	+	-
2	-	+
3	Standard	Standard
4	+	-
5	-	+

Tableau 4-8 : Conditions de passage des lots utilisés pour le test de la boucle de régulation

La Figure 4.31 présente les courbes d'endurance des deux cellules paire et impaire pour les différents lots utilisés. Les deux cellules présentent un comportement identique pour les différents lots dans l'évolution des tensions de seuil V_{TP} et V_{TE}. Ceci prouve que la boucle de régulation a pu compenser la variabilité de la largeur de la tranchée en modifiant la largeur de l'empilement mémoire pour maintenir contant la distance L_D. Cependant, nous remarquons une légère différence entre les cellules paire et impaire au niveau de la tension de seuil de programmation pour le lot 1. Cet écart est probablement dû à la dissymétrie des réticules de la tranchée et celui du transistor mémoire pendant le procédé de photolithographie.

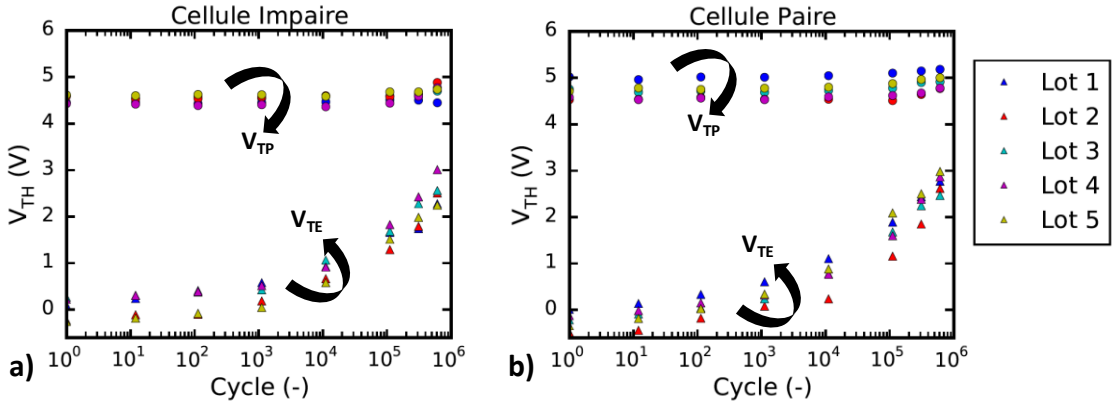


Figure 4.31 : Dispersion lot à lot des tensions V_{TE} et V_{TP} durant le test d'endurance pour les cellules paire et impaire

Comme observé précédemment sur les cellules eSTM, la dégradation de la fenêtre de programmation, de l'ordre de 54%, provient exclusivement de l'augmentation de la tension de seuil effacée V_{TE} passant de 0V à plus de 2V au fil des cycles. Malgré une tension de seuil V_{TP} légèrement plus élevée pour la cellule paire du lot 1, la dégradation de la fenêtre de programmation pour les deux cellules est identique (Figure 4.32).

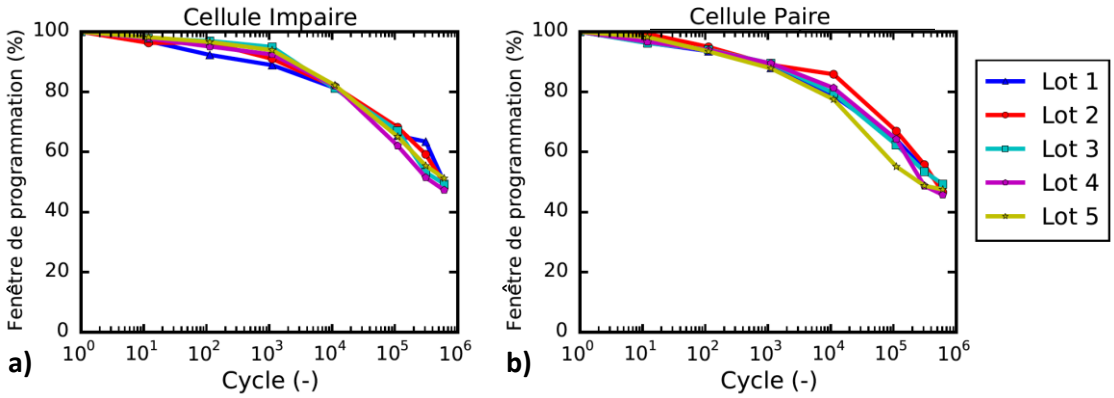


Figure 4.32 : Dispersion lot à lot de la fenêtre de programmation durant le test d'endurance pour les cellules paire et impaire

b) La variation intra-plaque de la fenêtre de programmation

Dans un premier temps, nous allons représenter l'évolution des tensions de seuil V_{TE} et V_{TP} minimale et maximale pour le lot de référence. La Figure 4.33 montre que la variation intra-plaque des tensions de seuil V_{TP} et V_{TE} est aux alentours de 200mV. Nous remarquons que les deux cellules présentent la même dispersion intra-plaque. Cette dispersion est identique à celle trouvée dans le paragraphe III.2.

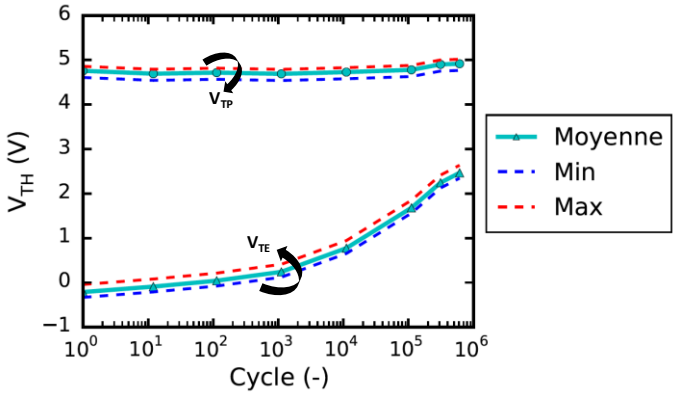


Figure 4.33 : Dispersion intra-plaque des tensions V_{TE} et V_{TP} durant le test d'endurance du lot3

Après avoir évalué la variation intra-plaque des tensions de seuil du lot de référence, nous allons maintenant faire une comparaison avec les autres lots. La Figure 4.34a représente la caractéristique d'endurance pour les lots avec la largeur de tranchée la plus élevée (lot 1&4). Il apparaît que la variation intra-plaque des tensions de seuil V_{TP} et V_{TE} est identique à celle du lot de référence. En ce qui concerne les lots avec la largeur de tranchée la plus faible (lot 2&5), le résultat est le même comparé au lot de référence (Figure 4.34b). Ces mesures d'endurance montrent que la boucle de régulation n'a pas d'impact négatif sur la variabilité intra-plaque des tensions de seuil V_{TE} et V_{TP}.

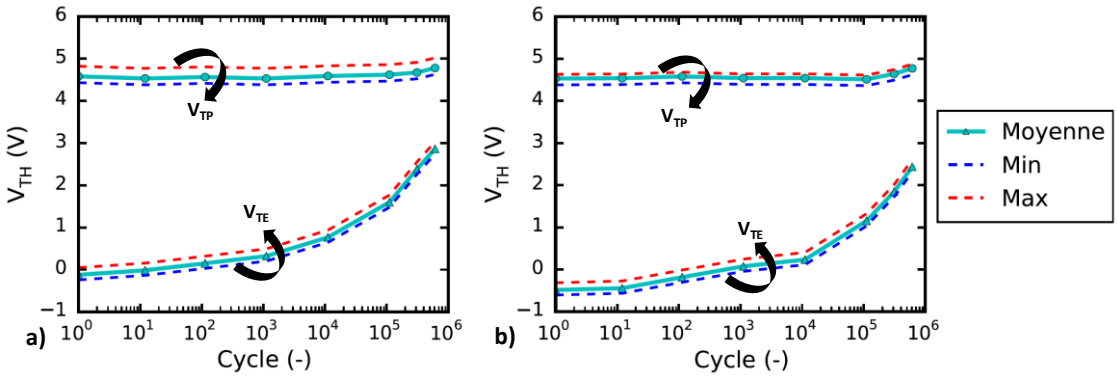


Figure 4.34 : Dispersion intra-plaque des tensions V_{TE} et V_{TP} durant le test d'endurance - a) Lot1&4 – b) Lot 2&5

Dans cette partie l'impact du nouveau procédé de gravure de l'empilement mémoire et de la boucle de régulation sur les performances de la cellule eSTM a été étudié. Les mesures électriques montrent que l'optimisation de la recette de gravure a amélioré les performances de la cellule eSTM, tandis que la mise en place de la boucle de régulation garantit un fonctionnement stable de notre dispositif.

IV. Conclusion

Dans la première partie de ce chapitre, nous avons amélioré le procédé de gravure de la grille mémoire. Il s'agit de remplacer l'argon comme élément chimique pour graver la couche du BARC par le tétrafluorure de carbone CF_4. Ce nouveau procédé a permis de réduire la rugosité de la grille mémoire de 70% par rapport à l'ancien procédé de gravure. Ce procédé a montré une diminution de 87% de la dispersion des tensions de seuil après effacement. Nous avons aussi montré qu'aucune retombée négative sur les autres paramètres électriques de la cellule mémoire n'est apparue.

L'amélioration de la rugosité de la grille mémoire nous a permis de travailler sur le déploiement d'une boucle de régulation afin de compenser la variabilité lot à lot de L_{TR}. La méthodologie suivie pour construire notre boucle de régulation est le fruit de l'état de l'art de recherches bibliographiques académiques et industrielles. Cependant, nous avons adapté cette méthodologie aux moyens disponibles et accessibles dans le cadre de notre étude.

Dans la deuxième partie de ce chapitre, nous avons mis en place une boucle de régulation *Feed-Forward* entre la gravure de la tranchée et celle de la grille mémoire. Le contrôleur s'appuie sur une mesure SEMCD de la largeur de la tranchée du transistor de sélection pour ajuster la largeur de la grille mémoire. L'objectif est de maintenir la distance entre la tranchée et l'empilement mémoire constante. Malgré le transfert de la technologie sur le site de Crolles, nous avons pu valider notre boucle de régulation avec un minimum de données. L'analyse des paramètres en ligne et électriques a montré une amélioration des performances. Pour répondre aux exigences de l'équipe d'automation de ST-Rousset, nous avons proposé des actions afin d'améliorer la robustesse de notre boucle de régulation. Malheureusement, le manque de temps et de ressources nous a empêché d'étudier l'implémentation de ces actions.

Pour finir nous avons évalué l'impact des optimisations du procédé de fabrication sur les performances électriques de l'eSTM. L'analyse de la variabilité intra-plaque des tensions V_{TE} et V_{TP} a montré que le nouveau procédé de fabrication améliore cette dernière. Nous avons pu montrer que la mise en place de la boucle de régulation permet de garantir la fiabilité de la cellule eSTM pour les différentes largeurs de tranchée.

Conclusion générale

Le travail réalisé au cours de cette thèse porte sur différents aspects de la variabilité du procédé de fabrication de la nouvelle architecture mémoire eSTM et de son impact sur les performances électriques. Plusieurs points ont été abordés lors de cette étude :

Dans le premier chapitre, nous avons étudié les différentes familles de mémoires non volatiles en ciblant plus particulièrement les mémoires à stockage de charges. Les méthodes de programmation et d'effacement ont été expliquées ainsi que les phénomènes parasites qui peuvent être rencontrés pendant leur durée de vie. Ensuite, nous avons présenté deux mémoires non-volatiles à stockage de charges, présentes dans la littérature, adressant les applications basse consommation : la mémoire 2T et la cellule Split Gate. Nous avons conclu ce chapitre avec l'introduction de la nouvelle architecture mémoire eSTM (*embedded Select Trench Memory*), inventée par STMicroelectronics, qui a pour objectif d'adresser les applications basse consommation.

Dans le deuxième chapitre, le procédé de fabrication de l'architecture eSTM et son principe de fonctionnement ont été présentés. La partie suivante consistait à caractériser électriquement le transistor vertical seul, ainsi que les cellules mémoires de part et d'autre du transistor de sélection. Durant ces tests, nous avons pu mettre en évidence des différences entre les cellules paire et impaire dues à la rugosité de la ligne de mot ou à un désalignement lors du procédé de fabrication. Nous avons aussi évalué le comportement électrique de cette cellule mémoire en appliquant différentes tensions (V_{GC}, V_D, V_{GS}) durant la programmation. La comparaison des performances électriques entre une cellule Flash standard et la cellule eSTM a montré que le courant de programmation est divisé par 10 pour la cellule eSTM, avec une amélioration après les tests d'endurance. Cette comparaison montre aussi des résultats très intéressants en termes de courant de programmation ce qui justifie la possibilité d'adresser les applications basse consommation.

Dans le troisième chapitre, nous avons essayé de mettre en place une méthodologie statistique qui permet d'identifier les causes à l'origine des variations des paramètres électriques. Cependant, le fait que l'architecture eSTM soit en cours de développement ne permet pas d'utiliser une méthodologie standard pour analyser la variabilité du procédé de fabrication. Pour contourner ce problème, nous avons conduit une campagne de mesures des étapes critiques du procédé de fabrication de l'eSTM pour évaluer la variabilité lot à lot, plaque à plaque, intra-champ et intra-die. Cette campagne nous a permis de proposer une Correction Optique de Proximité pour corriger la variabilité intra-die de la largeur de la zone d'active et celle de la tranchée du transistor de sélection. Nous avons aussi identifié la rugosité de la ligne mot et la variabilité temporelle de la largeur de la tranchée du transistor de sélection comme **sources majeures de la variabilité paramétrique**. Ces deux paramètres ont un impact sur un autre paramètre critique de la cellule eSTM : la distance entre la tranchée du transistor de sélection et l'empilement mémoire.

Dans le dernier chapitre, l'objectif était de réduire la rugosité de la grille mémoire et de contrôler la variabilité temporelle de la largeur de la tranchée du transistor de sélection. Après le remplacement de l'argon par le tétrafluorure de carbone CF_4 comme élément chimique pour

graver la couche du BARC, nous avons observé une réduction de 70% de la rugosité de ligne de mot. Ce nouveau procédé a montré une amélioration de 87% de la dispersion des tensions de seuil effacées des cellules sans avoir de retombée négative sur les autres paramètres électriques de la cellule mémoire.

L'amélioration de la rugosité de la grille mémoire nous a permis de travailler sur le déploiement d'une boucle de régulation *Feed-Forward* afin de compenser la variabilité lot à lot du paramètre L_{TR}. Pour développer cette boucle de régulation, nous avons adapté la méthodologie, issue de la littérature, aux moyens disponibles et accessibles dans le cadre de notre étude. Le contrôleur mis en place permet de compenser la dispersion de L_{TR} en ajustant la largeur de la grille mémoire. La mise en place de cette boucle de compensation a maintenu la distance entre le transistor de sélection et l'empilement mémoire constante. La dernière partie de chapitre montre que cette boucle de régulation garantit la fiabilité de la cellule eSTM malgré la variabilité de la largeur de la tranchée L_{TR}. En plus de ces résultats positifs, nous avons proposé des améliorations pour répondre aux exigences de l'équipe automation de STMicroelectronics de Rousset en matière de robustesse. Ces améliorations consistent à ajouter un filtre afin de garantir la fiabilité de la mesure en ligne.

La méthodologie présentée dans ce manuscrit pour trouver les sources de variabilité des paramètres électriques peut servir de référence pour d'autres travaux sur l'analyse de variabilité durant la phase de développement de nouvelles architectures. Les perspectives de nos travaux peuvent s'orienter vers trois directions : la première concerne une confirmation du bon fonctionnement de la boucle de régulation dans un environnement de production ; la deuxième perspective est de reproduire la même étude et d'adapter la boucle de régulation pour des nœuds technologiques de la cellule eSTM plus agressifs (40nm et 28nm). Il serait également intéressant d'ajouter la mesure d'alignement des réticules de la tranchée du transistor de sélection et celui de l'empilement mémoire au modèle prédictif du contrôleur. Cette action permettrait d'améliorer la robustesse de la boucle de régulation.

BIBLIOGRAPHIE

[1] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.," *IEEE Solid-State Circuits Newsl.*, vol. 20, no. 3, pp. 33–35, Sep. 2006.

[2] A. G. F. Dingwall and R. E. Strieker, "High density COS/MOS 1024 bit static RAM," in *1974 International Electron Devices Meeting (IEDM)*, 1974, pp. 101–103.

[3] A. C. Dumbri and W. Rosenzweig, "Static RAMs with microwatt data retention capability," *IEEE J. Solid-State Circuits*, vol. 15, no. 5, pp. 826–831, Oct. 1980.

[4] W. Sander, J. Early, and T. Longo, "A 4096 x 1 (I3L) bipolar dynamic RAM," in *1976 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, 1976, vol. XIX, pp. 182–183.

[5] D. Kahng and S. M. Sze, "A Floating Gate and Its Application to Memory Devices," *Bell Syst. Tech. J.*, vol. 46, no. 6, pp. 1288–1295, Jul. 1967.

[6] D. Frohman-Bentchkowsky, "MEMORY BEHAVIOR IN A FLOATING-GATE AVALANCHE-INJECTION MOS (FAMOS) STRUCTURE," *Appl. Phys. Lett.*, vol. 18, no. 8, p. 332, 1971.

[7] Y.-F. Chan, "A 4K CMOS erasable PROM," *IEEE J. Solid-State Circuits*, vol. 13, no. 5, pp. 677–680, Oct. 1978.

[8] T. Hagiwara, Y. Yatsuda, R. Kondo, S. Minami, T. Aoto, and Y. Itoh, "A 16 kbit electrically erasable PROM using n-channel Si-gate MNOS technology," *IEEE J. Solid-State Circuits*, vol. 15, no. 3, pp. 346–353, Jun. 1980.

[9] E. Harari, L. Schmitz, B. Troutman, and S. Wang, "A 256-bit nonvolatile static RAM," in *1978 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, 1978, vol. XXI, pp. 108–109.

[10] F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A new flash E^2PROM cell using triple polysilicon technology," in *1984 International Electron Devices Meeting*, 1984, vol. 30, pp. 464–467.

[11] G. Samachisa, C.-S. Chien-Sheng Su, Y.-S. Yu-Sheng Kao, G. Smarandoiu, T. Ting Wong, and C. Chenming Hu, "A 128K flash EEPROM using double polysilicon technology," in *1987 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, 1987, vol. XXX, pp. 76–77.

[12] P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, *Flash Memories*. Boston, MA: Springer US, 1999.

[13] R. H. Fowler and L. Nordheim, "Electron Emission in Intense Electric Fields," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 119, no. 781, pp. 173–181, May 1928.

[14] E. Takeda, Y. Nakagome, H. Kume, and S. Asai, "New hot-carrier injection and device degradation in submicron MOSFETs," *IEE Proc. I Solid State Electron Devices*, vol. 130, no. 3, p. 144, 1983.

[15] C. Chang, M.-S. Liang, C. Hu, and R. W. Brodersen, "Carrier tunneling related phenomena in thin oxide MOSFET's," in *1983 International Electron Devices Meeting*, 1983, vol. 29, pp. 194–197.

[16] T. H. Ning, "Hot-electron emission from silicon into silicon dioxide," *Solid. State. Electron.*, vol. 21, no. 1, pp. 273–282, Jan. 1978.

[17] E. Takeda, Y. Ohji, and H. Kume, "High field effects in MOSFETS," in *1985 International Electron Devices Meeting*, 1985, vol. 31, pp. 60–63.

[18] C. Chenming Hu, S. C. Simon C. Tam, F.-C. Fu-Chieh Hsu, P.-K. Ping-Keung Ko, T.-Y. Tung-Yi Chan, and K. W. Terrill, "Hot-Electron-Induced MOSFET Degradation - Model, Monitor, and Improvement," *IEEE J. Solid-State Circuits*, vol. 20, no. 1, pp. 295–305, Feb. 1985.

[19] S. Simon Tam, P.-K. Ping-Keung Ko, and C. Chenming Hu, "Lucky-electron model of channel hot-electron injection in MOSFET'S," *IEEE Trans. Electron Devices*, vol. 31, no. 9, pp. 1116–1125, Sep. 1984.

[20] H. Hidaka, "Evolution of embedded flash memory technology for MCU," in *2011 IEEE International Conference on IC Design & Technology*, 2011, pp. 1–4.

[21] K. Baker, "Embedded Nonvolatile Memories: A Key Enabler for Distributed Intelligence," in *2012 4th IEEE International Memory Workshop*, 2012, pp. 1–4.

[22] K. Naruke, S. Taguchi, and M. Wada, "Stress induced leakage current limiting to scale down EEPROM tunnel oxide thickness," in *Technical Digest., International Electron Devices Meeting*, 1988, pp. 424–427.

[23] L. D. Yau, "Simple I/V model for short-channel i.g.f.e.t.s in the triode region," *Electron. Lett.*, vol. 11, no. 2, p. 44, 1975.

[24] J. R. Brews, W. Fichtner, E. H. Nicollian, and S. M. Sze, "Generalized guide for MOSFET miniaturization," *IEEE Electron Device Lett.*, vol. 1, no. 1, pp. 2–4, Jan. 1980.

[25] W. Fichtner, E. N. Fuls, R. L. Johnston, T. T. Sheng, and R. K. Watts, "Experimental and theoretical characterization of submicron MOSFETs," in *1980 International Electron Devices Meeting*, 1980, vol. 26, pp. 24–27.

[26] M. Fukuma and M. Matsumura, "A simple model for short channel MOSFET's," *Proc. IEEE*, vol. 65, no. 8, pp. 1212–1213, 1977.

[27] J. R. Anderson, "Ferroelectric materials as storage elements for digital computers and switching systems," *Trans. Am. Inst. Electr. Eng. Part I Commun. Electron.*, vol. 71, no. 6, pp. 395–401, 1953.

[28] J.-H. Kim, D. J. Jung, S. K. Kang, Y. M. Kang, H. H. Kim, J. Y. Kang, E. S. Lee, W. W. Jung, H. J. Joo, J. Y. Jung, J. H. Park, H. Kim, D. Y. Choi, S. Y. Lee, H. S. Jeong, and K. Kim, "Manufacturing Technologies for a Highly Reliable, 0.34 μm2 Cell, 64 Mb, and 1T1C FRAM," in *2006 International Electron Devices Meeting*, 2006, pp. 1–4.

[29] A. Pohm, C. Sie, R. Uttecht, V. Kao, and O. Agrawal, "Chalcogenide glass bistable resistivity (Ovonic) memories," *IEEE Trans. Magn.*, vol. 6, no. 3, pp. 592–592, Sep. 1970.

[30] L. Goux, T. Gille, D. Tio Castro, G. A. M. Hurkx, J. G. Lisoni, R. Delhougne, D. J.

Gravesteijn, K. De Meyer, K. Attenborough, and D. J. Wouters, "Evidence of the Prominent Role of the Time-Under-Melt Parameter in the Reset Switching of Phase-Change Line Cells," in *2008 Joint Non-Volatile Semiconductor Memory Workshop and International Conference on Memory Technology and Design*, 2008, pp. 37–38.

[31] S. Lai, "Current status of the phase change memory and its future," in *IEEE International Electron Devices Meeting 2003*, 2003, p. 10.1.1-10.1.4.

[32] D. Ha and K. Kim, "Recent Advances in High Density Phase Change Memory (PRAM)," in *2007 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, 2007, pp. 1–4.

[33] R. Bez, "Chalcogenide PCM: a memory technology for next decade," in *2009 IEEE International Electron Devices Meeting (IEDM)*, 2009, pp. 1–4.

[34] M. Durlam, P. Naji, M. DeHerrera, S. Tehrani, G. Kerszykowski, and K. Kyler, "Nonvolatile RAM based on magnetic tunnel junction elements," in *2000 IEEE International Solid-State Circuits Conference. Digest of Technical Papers (Cat. No.00CH37056)*, 2000, pp. 130–131.

[35] T. Endoh, "STT-MRAM for low power systems," in *2015 International Symposium on VLSI Technology, Systems and Applications*, 2015, pp. 1–2.

[36] M. Durlam, Y. Chung, M. DeHerrera, B. . Engel, G. Grynkewich, B. Martino, B. Nguyen, J. Salter, P. Shah, and J. M. Slaughter, "MRAM Memory for Embedded and Stand Alone Systems," in *2007 IEEE International Conference on Integrated Circuit Design and Technology*, 2007, pp. 1–4.

[37] S. Seo, M. J. Lee, D. H. Seo, E. J. Jeoung, D.-S. Suh, Y. S. Joung, I. K. Yoo, I. R. Hwang, S. H. Kim, I. S. Byun, J.-S. Kim, J. S. Choi, and B. H. Park, "Reproducible resistance switching in polycrystalline NiO films," *Appl. Phys. Lett.*, vol. 85, no. 23, p. 5655, 2004.

[38] C. Rohde, B. J. Choi, D. S. Jeong, S. Choi, J.-S. Zhao, and C. S. Hwang, "Identification of a determining parameter for resistive switching of TiO[sub 2] thin films," *Appl. Phys. Lett.*, vol. 86, no. 26, p. 262907, 2005.

[39] T. W. Hickmott, "Low-Frequency Negative Resistance in Thin Anodic Oxide Films," *J. Appl. Phys.*, vol. 33, no. 9, p. 2669, 1962.

[40] B. Prince, *Vertical 3D memory technologies*. .

[41] G. Guoqiao Tao, H. Chauveau, D. Boter, E. van der Vegt, D. Dormans, and R. Verhaar, "Characterization and modeling of program/erase induced device degradation in 2T-FNFN-NOR flash memories," in *2008 15th International Symposium on the Physical and Failure Analysis of Integrated Circuits*, 2008, pp. 1–5.

[42] Y. K. Lee, J. H. Moon, Y. H. Kim, M.-J. Chun, S.-Y. Ha, S. Choi, H. Yoo, H. Jeon, J. Yu, J.-U. Han, E. Jung, and C. Chung, "2T-FN eNVM with 90 nm Logic Process for Smart Card," in *2008 Joint Non-Volatile Semiconductor Memory Workshop and International Conference on Memory Technology and Design*, 2008, pp. 26–27.

[43] M. van Duuren, R. van Schaijk, M. Slotboom, P. Tello, P. Goarin, N. Akil, F. Neuilly, Z. Rittersma, and A. Huerta, "Performance and Reliability of 2-Transistor FN/FN Flash Arrays with Hafnium Based High-K Inter-Poly Dielectrics for Embedded NVM," in *2006 21st IEEE Non-Volatile Semiconductor Memory Workshop*, 2006, pp. 48–49.

[44] S.-R. Kim, K. J. Han, K.-S. Lee, R. Li, J. Wolfman, T.-H. Kim, P. Liu, H. Kim, P.-Y. Lee, Y. Wang, Y. Jia, F. Dhaoui, F. Hawley, and H.-C. Tseng, "High performance 65nm 2T-embedded Flash memory for high reliability SOC applications," in *2010 IEEE International Memory Workshop*, 2010, pp. 1–3.

[45] S. Kianian, A. Levi, D. Lee, and Y.-W. Yaw-Wen Hu, "A novel 3 volts-only, small sector erase, high density flash E/sup 2/PROM," in *Proceedings of 1994 VLSI Technology Symposium*, 1994, pp. 71–72.

[46] Y. Tkachev, X. Liu, and A. Kotov, "Floating-Gate Corner-Enhanced Poly-to-Poly Tunneling in Split-Gate Flash Memory Cells," *IEEE Trans. Electron Devices*, vol. 59, no. 1, pp. 5–11, Jan. 2012.

[47] A. T. Tilke, L. Pescini, M. Bauer, M. Stiftinger, R. Kakoschke, D. Shum, N. Chan, S. Kim, V. Hecht, and K. J. Han, "Highly Scalable Embedded Flash Memory With Deep Trench Isolation and Novel Buried Bitline Integration for the 90-nm Node and Beyond," *IEEE Trans. Electron Devices*, vol. 54, no. 7, pp. 1681–1688, Jul. 2007.

[48] A. Watson and S. H. Voldman, "The effect of deep trench and sub-collector on the latchup robustness in BiCMOS silicon germanium technology," in *Bipolar/BiCMOS Circuits and Technology, 2004. Proceedings of the 2004 Meeting*, 2004, pp. 172–175.

[49] H.-L. Tu, I.-S. Chen, P.-C. Yeh, and H.-K. Chiou, "High Performance Spiral Inductor on Deep-Trench-Mesh Silicon Substrate," *IEEE Microw. Wirel. Components Lett.*, vol. 16, no. 12, pp. 654–656, Dec. 2006.

[50] H. Sunami, "Development of three-dimensional MOS structures from trench-capacitor DRAM cell to pillar-type transistor," in *2008 9th International Conference on Solid-State and Integrated-Circuit Technology*, 2008, pp. 853–856.

[51] D. Lee, F. Tsui, J.-W. Jeng-Wei Yang, F. Feng Gao, W.-J. Wen-Juei Lu, Y. Yeeheng Lee, C.-T. Chi-Tsai Chen, V. Huang, P.-Y. Pin-Yao Wang, M. H. Liu, H. C. Hsu, S. Chang, S. Y. Chang, H. Van Tran, J. Frayer, B. Yaw-Wen Hu, B. Yeh, and B. Chen, "Vertical floating-gate 4.5F2 split-gate NOR flash memory at 110nm node," in *Digest of Technical Papers. 2004 Symposium on VLSI Technology, 2004.*, 2004, pp. 72–73.

[52] H. Van Tran, A. Ly, V. Sarin, S. T. Nguyen, H. Q. Nguyen, L. Hoang, H. B. Kim, I. Nojima, D. Lee, and B. Chen, "An experimental 1Mb 0.11 um 4.5F2 1.8Volt multilevel vertical split gate source side injection test vehicle for giga-bit density NOR flash memory," *2005 IEEE Asian Solid-State Circuits Conf. ASSCC 2005*, pp. 125–128, 2006.

[53] B. Chen, "Highly Reliable SuperFlash Embedded Memory Scaling for Low Power SoC," in *2007 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, 2007, pp. 1–2.

[54] N. Sullivan, L. L. Raja, R. J. Kee, Y. Yokota, and M. Williams, "Exploring ISSG process space [Si oxidation]," in *9th International Conference on Advanced Thermal Processing of Semiconductors, RTP 2001*, pp. 95–110.

[55] J. Yohan, "Étude des fluctuations locales des transistors MOS destinés aux applications analogiques," 2011.

[56] J.Bartoli, "Développement et caractérisation d'architectures mémoires non volatiles pour les applications basse consommation," 2014.

[57] B. De Salvo, G. Ghibaudo, G. Pananakakis, G. Reimbold, F. Mondond, B. Guillaumot, and P. Candelier, "Experimental and theoretical investigation of nonvolatile memory data-retention," *IEEE Trans. Electron Devices*, vol. 46, no. 7, pp. 1518–1524, Jul. 1999.

[58] J. (Joe E. . Brewer and M. Gill, *Nonvolatile memory technologies with emphasis on Flash : a comprehensive guide to understanding and using NVM devices*. IEEE Press, 2008.

[59] D. Shum, J. R. Power, R. Ullmann, E. Suryaputra, K. Ho, J. Hsiao, C. H. Tan, W. Langheinrich, C. Bukethal, V. Pissors, G. Tempel, M. Rohrich, A. Gratz, A. Iserhagen, E. O. Andersen, S. Paprotta, W. Dickenscheid, R. Strenz, R. Duschl, T. Kern, C. T. Hsieh, C. M. Huang, C. W. Ho, H. H. Kuo, C. W. Hung, Y. T. Lin, and L. C. Tran, "Highly Reliable Flash Memory with Self-Aligned Split-Gate Cell Embedded into High Performance 65nm CMOS for Automotive & Smartcard Applications," in *2012 4th IEEE International Memory Workshop*, 2012, pp. 1–4.

[60] S. Yamada, Y. Hiura, T. Yamane, K. Amemiya, Y. Ohshima, and K. Yoshikawa, "Degradation mechanism of flash EEPROM programming after program/erase cycles," in *Proceedings of IEEE International Electron Devices Meeting*, pp. 23–26.

[61] S. S. Chung, C. M. Yih, S. M. Cheng, and M. S. Liang, "A New Oxide Damage Characterization Technique For Evaluating Hot Carrier Reliability Of Flash Memory Cell After P/E Cycles," in *Symposium on VLSI Technology*, 1997, pp. 111–112.

[62] S. S. Chung, Cherng-Ming Yih, Shui-Ming Cheng, and Mong-Song Liang, "A new technique for hot carrier reliability evaluations of flash memory cell after long-term program/erase cycles," *IEEE Trans. Electron Devices*, vol. 46, no. 9, pp. 1883–1889, 1999.

[63] P. Cappelletti, R. Bez, D. Cantarelli, and L. Fratin, "Failure mechanisms of flash cell in program/erase cycling," in *Proceedings of 1994 IEEE International Electron Devices Meeting*, pp. 291–294.

[64] G. JUST, "Caractérisation et modélisation des mémoires Flash embarquées destinées aux applications faible consommation et à forte contrainte de fiabilité."

[65] V. Della Marca, J. Postel-Pellerin, G. Just, P. Canet, and J.-L. Ogier, "Impact of endurance degradation on the programming efficiency and the energy consumption of NOR flash memories," *Microelectron. Reliab.*, vol. 54, pp. 2262–2265, 2014.

[66] B. E. Stine, D. S. Boning, J. E. Chung, D. a Bell, and E. Equi, "SPIE Symposium on Microelectronic Manufacturing , SPIE Vol. 2874, p. 27, Austin, TX, Oct. 1996.," vol. 2874, 1996.

[67] et S. H. Joost Van Herk, Jean De Caunes, Francois Pasqualini, *Guidline to start a R2R control loop*. 2005.

[68] C. J. Raymond, S. S. H. Naqvi, and J. R. McNeil, "Scatterometry for CD measurements of etched structures," 1996, pp. 720–728.

[69] J. C. Vickerman and D. (David) Briggs, *TOF-SIMS : materials analysis by mass spectrometry*. .

[70] Young-Bog Park and D. K. Schroder, "Degradation of thin tunnel gate oxide under constant Fowler-Nordheim current stress for a flash EEPROM," *IEEE Trans. Electron Devices*, vol. 45, no. 6, pp. 1361–1368, Jun. 1998.

[71] U. Buttgereit, R. Birkner, T. Scheruebl, S. de Putter, B. Kastrup, and J. Finders, "Reducing

the impact of reticle CD-non-uniformity of multiple structures by dose corrections based on aerial image measurements," 2010, p. 76380D.

[72] H.-J. Kwon, D.-S. Min, P.-J. Jang, B.-S. Chang, B.-Y. Choi, and S.-H. Jeong, "Loading effect parameters of dry etcher system and their analysis in mask-to-mask loading and within-mask loading," 2002, p. 79.

[73] J. Kang, J. Kim, S. Jung, H. Kim, and K. Kim, "Modified optical proximity correction model to compensate pattern density induced optical proximity effect," 2005, p. 1220.

[74] B. E. Stine, D. S. Boning, and J. E. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices," *IEEE Trans. Semicond. Manuf.*, vol. 10, no. 1, pp. 24–41, 1997.

[75] María Rízquez, "Characterization and optimization of high density plasma etching processes for advanced memories application," 2016.

[76] G. Roy, A. Ghetti, A. Benvenuti, A. Erlebach, and A. Asenov, "Comparative simulation study of the different sources of statistical variability in contemporary floating-gate nonvolatile memory," *IEEE Trans. Electron Devices*, vol. 58, no. 12, pp. 4155–4163, 2011.

[77] E. A. Agharben, A. Roussy, E. A. Agharben, M. Bocquet, E. A. Agharben, M. Bileci, S. Begouin, and A. Marchadier, "Critical sensitivity of flash gate dimension spread on electrical performances for advanced embedded memory," in *2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2015, pp. 401–404.

[78] X. Jiang, M. Li, R. Wang, J. Chen, and R. Huang, "Investigations on the correlation between line-edge-roughness (LER) and line-width-roughness (LWR) in nanoscale CMOS technology," in *2012 IEEE 11th International Conference on Solid-State and Integrated Circuit Technology*, 2012, pp. 1–3.

[79] F. Zhao, L. Zhang, Q. Wang, and Z. Jiang, "Impact of line edge roughness and linewidth roughness on critical dimension variation," in *2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, 2012, vol. 3, pp. 475–479.

[80] Z. Jiang, F. Zhao, W. Jing, P. D. Prewett, and K. Jiang, "Characterization of line edge roughness and line width roughness of nano-scale typical structures," in *2009 4th IEEE International Conference on Nano/Micro Engineered and Molecular Systems*, 2009, pp. 299–303.

[81] G. Ayal, D. Andelman, and Y. Cohen, "Analytical model for ArF photoresist shrinkage under scanning electron microscopy inspection," *J. Vac. Sci. Technol. B Microelectron. Nanom. Struct.*, vol. 27, no. 4, p. 1976, 2009.

[82] J. Lowes, V. Pham, J. Meador, C. Stroud, F. Rosas, R.-M. L. Mercado, and M. Slezak, "Advantages of BARC and photoresist matching for 193-nm photosensitive BARC applications," 2010, p. 76390K.

[83] R. Huang and M. Weigand, "Plasma etch properties of organic BARCs," 2008, p. 69232G.

[84] S. Ruegsegger, A. Wagner, J. S. Freudenberg, and D. S. Grimard, "Feedforward control for reduced run-to-run variation in microelectronics manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 12, no. 4, pp. 493–502, 1999.

[85] N. Jedidi, P. Sallagoity, A. Roussy, and S. Dauzere-Peres, "Feedforward Run-to-Run Control for Reduced Parametric Transistor Variation in CMOS Logic 0.13 μm

Technology," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 2, pp. 273–279, May 2011.

[86] R. Good and S. J. Qin, "Stability analysis of double EWMA run-to-run control with metrology delay," in *Proceedings of the 2002 American Control Conference (IEEE Cat. No.CH37301)*, 2002, pp. 2156–2161 vol.3.

[87] D. C. Montgomery, *Design and analysis of experiments*. John Wiley & Sons, Inc, 2013.

[88] "Design of Experiments," in *Analytic Methods for Design Practice*, London: Springer London, 2007, pp. 309–391.

[89] K. Faron, M. Freeland, O. Krogh, S. Patel, and G. Raghavendra, "Multivariable versus univariable APC," 2004, p. 18.

[90] E. Lenderink and P. Stehouwer, "Optimization, sensitivity analysis, and robust design using response surface modeling," 2008, p. 710302.

[91] S. Adivikolanu and E. Zafiriou, "Extensions and performance/robustness tradeoffs of the EWMA run-to-run controller by using the internal model control structure," *IEEE Trans. Electron. Packag. Manuf.*, vol. 23, no. 1, pp. 56–68, Jan. 2000.

[92] A. Badr, "General Introduction to Design of Experiments (DOE)," in *Wide Spectra of Quality Control*, InTech, 2011.

[93] American Society for Quality Control., *Journal of quality technology*. American Society for Quality Control.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT : *2017LYSEM013*

El Amine AGHARBEN

Optimisation et réduction de la variabilité d'une nouvelle architecture mémoire non volatile ultra basse consommation

Spécialité : Microélectronique

Mots clefs : Mémoire non volatile, Analyse de variabilité, Boucle de régulation, …

Résumé :

Le marché mondial des semi-conducteurs connait une croissance continue due à l'essor de l'électronique grand public et entraîne dans son sillage le marché des mémoires non volatiles. L'importance de ces produits mémoires est accentuée depuis le début des années 2000 par la mise sur le marché de produits nomades tels que les smartphones ou plus récemment les produits de l'internet des objets. De par leurs performances et leur fiabilité, la technologie Flash constitue, à l'heure actuelle, la référence en matière de mémoire non volatile. Cependant, le coût élevé des équipements en microélectronique rend impossible leur amortissement sur une génération technologique. Ceci incite l'industriel à adapter des équipements d'ancienne génération à des procédés de fabrication plus exigeants. Cette stratégie n'est pas sans conséquence sur la dispersion des caractéristiques physiques (dimension géométrique, épaisseur…) et électriques (courant, tension…) des dispositifs. Dans ce contexte, le sujet de ma thèse est d'optimiser et de réduire la variabilité d'une nouvelle architecture mémoire non volatile ultra basse consommation.

Cette étude vise à poursuivre les travaux entamés par STMicroelectronics sur le développement, l'étude et la mise en œuvre de boucles de contrôle de type Run-to-Run (R2R) sur une nouvelle cellule mémoire ultra basse consommation. Afin d'assurer la mise en place d'une régulation pertinente, il est indispensable de pouvoir simuler l'influence des étapes du procédé de fabrication sur le comportement électrique des cellules en s'appuyant sur l'utilisation d'outils statistiques ainsi que sur une caractérisation électrique pointue.

École Nationale Supérieure des Mines
de Saint-Étienne

NNT: *2017LYSEM013*

El Amine AGHARBEN

Optimization and reduction of the variability of a new nonvolatile memory architecture ultra-low power consumption

Specialty: Microelectronics

Keywords: Nonvolatile memory, ANOVA, Run-to-Run, …

Abstract:

The global semiconductor market is experiencing steady growth due to the development of consumer electronics and the wake of the non-volatile memory market. The importance of these memory products has been accentuated since the beginning of the 2000s by the introduction of nomadic products such as smartphones or, more recently, the Internet of things. Because of their performance and reliability, Flash technology is currently the standard for non-volatile memory. However, the high cost of microelectronic equipment makes it impossible to depreciate them on a technological generation. This encourages industry to adapt equipment from an older generation to more demanding manufacturing processes. This strategy is not without consequence on the spread of the physical characteristics (geometric dimension, thickness ...) and electrical (current, voltage ...) of the devices. In this context, the subject of my thesis is "Optimization and reduction of the variability of a new architecture ultra-low power non-volatile memory".

This study aims to continue the work begun by STMicroelectronics on the improvement, study and implementation of Run-to-Run (R2R) control loops on a new ultra-low power memory cell. In order to ensure the implementation of a relevant regulation, it is essential to be able to simulate the process manufacturing influence on the electrical behavior of the cells, using statistical tools as well as the electric characterization.

PUBLICATIONS

E. AGHARBEN, A. ROUSSY, M. BILECI, M. BOCQUET. "Critical sensitivity of Flash gate dimension spread on electrical performances for advanced embedded memory" In: (ASMC2015), 26th SEMI Advanced Semiconductor Manufacturing Conference, Mai 2015, Saratoga Springs, New York, US (Oral)

E. AGHARBEN, A. ROUSSY, M. BILECI, M. BOCQUET. "Flash gate optimized process and integration for electrical performances requirement on advanced embedded memory" In: (ISSM2016), International Symposium on Semiconductor, Dec 2016, Tokyo, Japan