METHODOLOGIE DE DIAGNOSTIC ET TECHNIQUES DE TEST POUR LES MEMOIRES NON VOLATILES DE TYPE EEPROM

Hassen Aziza

▶ **To cite this version:**

Hassen Aziza. METHODOLOGIE DE DIAGNOSTIC ET TECHNIQUES DE TEST POUR LES MEMOIRES NON VOLATILES DE TYPE EEPROM. Micro et nanotechnologies/Microélectronique. Aix-Marseille Université (AMU), 2004. Français. NNT : . tel-03504841

HAL Id: tel-03504841
https://hal.science/tel-03504841

Submitted on 29 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE DE PROVENCE AIX-MARSEILLE I
U.F.R. SCIENCES DE LA MATIERE

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE DE PROVENCE

Ecole doctorale : Physique, Modélisation et Sciences pour l'Ingénieur
Spécialité : Phénomènes Hors Equilibre, Micro et Nanoélectronique

présentée et soutenue publiquement
par

HASSEN AZIZA

Ingénieur ICF

METHODOLOGIE DE DIAGNOSTIC ET TECHNIQUES DE TEST POUR LES MEMOIRES NON VOLATILES DE TYPE EEPROM

Directrice de thèse : Annie PEREZ

Soutenue le 30 Novembre 2004 devant la commission d'examen :

Rapporteurs	Chantal Robach	Professeur, ESISAR Valence
	Michel Renovell	Directeur de Recherche CNRS, Montpellier II
Examinateurs	Annie Pérez	Maître de conférences, Université de Provence
	Rachid Bouchakour	Professeur, Université de Provence
	Didier Née	Chef de service, STMicroelectronics, Rousset
	Jean-Michel Portal	Maître de conférences, Université de Provence

REMERCIEMENTS

Je tiens en premier lieu à remercier Michel LANOO, directeur du Laboratoire de Matériaux et de Microélectronique de Provence (L2MP), mon laboratoire d'attache, ainsi que Rachid BOUCHAKOUR pour m'avoir accueilli au sein l'équipe microélectronique de ce laboratoire.

Je tiens à remercier chaleureusement Jean Michel PORTAL et Didier NEE pour m'avoir offert l'opportunité d'effectuer cette thèse. J'ai beaucoup profité de la rigueur scientifique et de l'esprit d'initiative de Jean Michel, notamment dans le domaine du test de circuits. Quand à Didier, sa longue expérience dans le semiconducteur et ses compétences dans les procédés de fabrication m'ont permis de m'investir sereinement dans mon travail de thèse. Ma reconnaissance va également à ma directrice de thèse, Annie PEREZ, pour la confiance qu'elle m'a accordée et sans qui l'accomplissement de cette thèse n'aurait pas été possible.

Je remercie vivement Madame Chantal ROBACH, professeur à l'ESISAR Valence, et monsieur Michel RENOVELL, directeur de Recherche à l'Université de Montpellier II, pour l'honneur qu'ils m'ont fait et pour l'intérêt qu'ils ont porté à mes travaux en acceptant d'être les rapporteurs de ce mémoire de thèse.

Je remercie également toute l'équipe du L2MP pour sa son soutient et sa bonne humeur, ce qui m'a permis d'évoluer dans les meilleures conditions. Je voudrais aussi exprimer mon amitié sincère à mes collègues doctorants et à toutes les personnes du laboratoire que j'ai côtoyées avec plaisir.

Mes remerciements vont aussi à toute l'équipe du CRD Rousset 8" (STMicroelectronics) et tout particulièrement à Julien VAST pour sa disponibilité et ses idées. Je souhaite également exprimer toute ma gratitude à toutes les personnes qui m'ont fait partager leur connaissance et leur savoir-faire au cours de ces années de thèse, en particulier Bertrand DELSUC, Laurent LOPEZ, Nicolas DEMANGE, Bruno CASADEI, Virginie BIDAL, Pascal FORNARA. Je tiens également à remercier Lionel FORLI et Bertrand SAILLET pour leur esprit d'équipe.

Je remercie les personnes du Centre de Recherche et Développement de ST Agrate (Italie) qui m'ont apporté des informations précieuses dans le domaine du test des mémoires non volatiles. Je pense tout particulièrement Daniele CANTARELI et Giovana DALLALIBERA.

Enfin, je tenais à exprimer des remerciements particuliers à l'Association Nationale de la Recherche Technique (ANRT) qui a permis la réalisation de ce projet de thèse.

Table des Matières

Table des symboles et des abréviations

Nom	Unité	Description
A	$A\ V^{-2}$	Coefficient Fowler-Nordheim
B	$V\ m^{-1}$	Coefficient Fowler-Nordheim
C_{pp}	F	Capacité de l'oxyde interpoly ONO
C_{ox}	F	Capacité d'oxyde
C_T	F	Capacité totale
C_{tun}	F	Capacité de l'oxyde tunnel
D_0	Déf./cm^2	Défauts par cm^2 et par niveau de masquage
D_{it}	$eV^{-1}\ m^{-2}$	Densité des états d'interface
E	$V\ m^{-1}$	Champ électrique
E_f	J	Energie du niveau de Fermi
E_{fm}	J	Potentiel électrochimique (ou niveau de Fermi) du métal
E_{fn}	J	Energie du quasi niveau de Fermi pour les électrons
E_{fp}	J	Energie du quasi niveau de Fermi pour les trous
E_g	J	Largeur de la bande interdite du semi-conducteur
E_i	j	Niveau d'énergie intrinsèque loin de l'interface
E_{ox}	$V\ m^{-1}$	Champ électrique dans l'oxyde
E_{tun}	$V\ m^{-1}$	Champ électrique dans l'oxyde tunnel
E_V	J	Energie du niveau de la bande de valence
g_m	$A\ V^{-1}$	Transconductance
h	J s	Constante de Planck ($h = 6{,}62\times10^{-34}$ J s)
ħ	J s	Constante de Planck réduite ($\hbar = h/2\pi$)
I_B	A	Courant dans le substrat
I_{DS}	A	Courant passant entre le drain et la source d'un transistor MOS
I_G	A	Courant dans la grille
I_{ON}	A	Courant I_{DS} traversant le canal d'un transistor dans l'état passant
I_{OFF}	A	Courant I_{DS} traversant le canal d'un transistor dans l'état bloqué
I_{ref}	A	Courant de lecture de référence
I_T	A	Courant de seuil
$I_{Tefface}$	A	Courant de seuil effacé du transistor mémoire
I_{Tecrit}	A	Courant de seuil écrit du transistor mémoire
$I_{Tvierge}$	A	Courant de seuil vierge du transistor mémoire
J_{FN}	$A\ m^{-1}$	Densité de courant Fowler-Nordheim
k	$J\ K^{-1}$	Constante de Boltzmann ($k = 1{,}38\times10^{-23}$ J K^{-1})
L	m	Longueur d'une capacité ou d'un transistor
L_{eff}	m	Longueur effective du canal du transistor mémoire
L_{tun}	m	Longueur de la fenêtre d'oxyde de tunnel du transistor mémoire
m_e	kg	masse de l'électron ($m_e = 9{,}1\times10^{-31}$ kg)
m^*_{ox}	kg	masse des électrons dans l'oxyde
m^*_{Si}	kg	masse des électrons dans le silicium

Nom	**Unité**	**Description**
N_A	m^{-3}	Concentration en atomes accepteurs
n	m^{-3}	Concentration d'électrons
n_0	m^{-3}	Concentration de trous libres dans le substrat loin de l'interface
n_i	m^{-3}	Concentration intrinsèque des porteurs
p	m^{-3}	Concentration de trous
p_0	m^{-3}	Concentration d'électrons libres dans le substrat loin de l'interface
q	C	Valeur absolue de la charge de l'électron ($q = 1{,}602\times10^{-19}$ C)
Q_{BD}	C m^{-2}	Charge de claquage
Q_{dep}	C m^{-2}	Charge de la couche de déplétion
Q_{FG}	C m^{-2}	Charge présente dans la grille flottante
Q_{FG0}	C m^{-2}	Charge initiale de grille flottante
Q_g	C m^{-2}	Charge de grille de contrôle
Q_{inv}	C m^{-2}	Charge de la couche d'inversion
Q_{it}	C m^{-2}	Charge des états d'interface
Q_{ox}	C m^{-2}	Charge dans l'oxyde
Q_{SC}	C m^{-2}	Charge dans le semi-conducteur
S_{pp}	m^2	Superficie de l'oxyde inter polysilicium
S_{tun}	m^2	Superficie de la fenêtre tunnel
T	K	Température
T_{ox}	m	Epaisseur d'oxyde
T_{prog}	s	Durée de l'impulsion de programmation
T_{pp}	m	Epaisseur de l'oxyde inter polysilicium du transistor mémoire
T_{tun}	m	Epaisseur de l'oxyde tunnel
V_{BL}	V	Tension appliquée sur la Bit Line
V_D	V	Tension appliquée sur le drain
V_{fb}	V	Tension de bandes plates
V_{FG}	V	Potentiel de grille flottante
V_{GC}	V	Tension appliquée sur la grille de contrôle
V_{max}	m s^{-1}	Vitesse de saturation des porteurs (électrons ~ 100.000 m s^{-1})
V_{pp}	V	Tension de l'impulsion de programmation
V_{ref}	V	Tension de lecture de référence
V_S	V	Tension appliquée sur la source
V_{Select}	V	Tension appliquée sur la grille du transistor de sélection
V_T	V	Tension de seuil d'un transistor
$V_{Tefface}$	V	Tension de seuil effacée du transistor mémoire
V_{Tecrit}	V	Tension de seuil écrite du transistor mémoire
$V_{Tvierge}$	V	Tension de seuil vierge du transistor mémoire
W	m	Largeur du transistor
W_{eff}	m	Largeur effective dans le canal du transistor mémoire
W_{tun}	m	Largeur de la zone d'injection tunnel
X_j	m	Profondeur de jonction
α_d		Coefficient de couplage en écriture
β	C J^{-1}	$\beta = q / kT$
ε_0	F m^{-1}	Permittivité du vide ($\varepsilon_0 = 8{,}85\times10^{-12}$ F m^{-1})
ε_{ox}	F m^{-1}	Constante diélectrique de l'oxyde SiO_2 ($\varepsilon_{ox} = 3{,}9$)
ε_S	F m^{-1}	Constante diélectrique du Silicium ($\varepsilon_S = 11{,}9$)
Φ_0	eV	Hauteur de barrière à l'interface injectante
α_g		Coefficient de couplage en effacement
λ	m	Longueur d'onde
Ψ_S	V	Potentiel de surface dans le semi-conducteur
Ψ_b	V	Potentiel de substrat, loin de l'interface

Nom	Unité	Description
μ_0	$m^2V^{-1}s^{-1}$	Mobilité des électrons dans le canal à faible champ électrique
BISD		Built In Self Diagnosis
BL		Bit Line
BPSG		Boron and Phosphorus doped Silicon Glass
BSIM		Berkeley Short channel IGFET Model
CAM		Content Adressable Memory
CAST		Cell Array Stress Test
CDMA		Current Direct Memory Access
CMOS		Complementary Metal Oxide Semiconductor
CVD		Chemical Vapor Deposition
DCE		Dichloroéthylène
DIBL		Drain Induced Barrier Lowering
DMA		Direct Memory Access
DRAM		Dynamic Random Access Memory
EAPROM		Electrically Alterable Programmable Read Only Memory
ECC		Error Correction Code
EDC		Error Detection Code
EEPROM		Electrically Erasable and Programmable Read Only Memory
EPROM		Electrically Programmable Read Only Memory
ETOX		Eprom Tunnel OXide
EWS		Electrical Wafer Sort
FLOTOX		FLOating gate Thin OXide
FN		Fowler-Nordheim
FRAM		Ferroelectric Random Access memory
GF		Grille Flottante
GIDL		Gate Induced Drain Leakage
GS		Ground Select
HV(OX)		High Voltage (OXide)
IMD		Inter-Metal Dielectric
I^2C		Inter Integrated Circuit
LDD		Low Drain Diffusion
LPCVD		Low-Pressure Chemical Vapor Deposition
LV(OX)		Low Voltage (OXide)
MIM		Metal-Isolator-Metal
MIMIS		Metal-Insulator-Metal-Insulator-Semiconductor
MM9		Mos Model 9
MNOS		Metal-Nitride-Oxide-Semiconductor
MOS		Metal-Oxide-Semiconductor
NOVRAM		NOn Volatile RAM
Nwell		Caisson de type N
ONO		Oxide-Nitride-Oxide (SiO2/Si3N4/SiO2)
OUM		Ovonic Unified Memory
PMD		Pre-Metal Dielectric
Pwell		Caisson de type P
RAM		Random Access Memory
ROM		Read Only Memory
RTA		Rapid Thermal Annealing
SAS		Self-Aligned Source
SEM		Scanning Electronic Microscopy
SILC		Stress Induced Leakage Current
SIMOS		Stacked gate Injection MOS
SL		Select Line

Nom	**Unité**	**Description**
SNOS		Silicon-Nitride-Oxide-Semiconductor
SONOS		Silicon-Oxide-Nitride-Oxide-Semiconductor
SPI		Serial Peripheral Interface
SPICE		Simulation Program with Integrated Circuit Emphasis
SRAM		Static Random Access Memory
TEM		Transmission Electronic Microscope
TEOS		TétraEthylOrtholoSilicate
TPFG		Textured Poly Floating Gate
USG		Undoped Silicon Glass
UV		Ultra-Violet
WKB		Wentzel, Kramers et Brillouin
WL		Word Line
ZCE		Zone de Charge d'Espace

Introduction générale

Le développement de la microélectronique depuis ces 30 dernières années est véritablement spectaculaire. Ce succès résulte en grande partie d'un savoir-faire et d'une maîtrise technologique de plus en plus poussés de l'élément fondamental de la microélectronique : le silicium. En l'espace de quelques années, le formidable essor de la microélectronique a conduit au développement des technologies de l'information qui ont envahi notre quotidien (informatique, télécommunications, électronique grand public…). Cette évolution rapide du marché des semiconducteurs repose sur des produits porteurs telles que les mémoires non volatiles de type EEPROM.
En effet, chaque application électronique, du domaine des communications ou de l'informatique, nécessite la possibilité de mémoriser des informations, que ce soit des instructions (program storage) ou des données de référence (data storage). Les mémoires non volatiles de type EEPROM permettent de garder l'information durant plus de 10 ans sans alimentation électrique, et offrent l'avantage d'être programmables et effaçables électriquement. Ces caractéristiques répondent à deux exigences majeures des systèmes les plus avancés : d'une part, la non volatilité qui permet de stocker des informations dans les applications portables, d'autre part la programmation électrique dans l'application qui permet de modifier et donc de faire évoluer le système. Les EEPROM, dont la caractéristique majeure est la fiabilité, constituent des composants stratégiques dans les systèmes complexes. Leur taille ne cesse d'augmenter et les applications qui les utilisent sont de plus en plus nombreuses. Aussi, la technologie EEPROM joue un rôle crucial dans le développement du marché de la carte à puce.
Cependant, ces types de mémoires sont de plus en plus confrontés à des problèmes de fiabilité dus essentiellement à la réduction des dimensions de la cellule mémoire élémentaire, réduction imposée par un accroissement de la densité d'intégration au niveau de la matrice de cellules mémoires. Dans ce contexte, le bon fonctionnement du dispositif est strictement lié à une meilleure compréhension des mécanismes de défaillance affectant aussi bien la cellule mémoire élémentaire que la matrice de cellules EEPROM. Par conséquent, la mise en place de solutions de test permettant le diagnostic de défauts dus au processus de fabrication, ou à des problèmes de conception, est primordiale.
L'objectif de cette thèse est de mettre en place une méthodologie de diagnostic de défauts spécifique aux mémoires EEPROM pouvant servir à identifier l'origine d'une défaillance. Le but final étant une amélioration du principal critère qui caractérise les performances d'une unité de production : le rendement de fabrication.
Le paramètre électrique le plus informatif dans le cas d'une cellule EEPROM est la tension de seuil V_T. Cette méthodologie de diagnostic se basera donc sur l'extraction de signatures électriques pertinentes, comme les tensions de seuil, sur le produit EEPROM.

Le **premier chapitre** se compose de deux parties. Une première partie, consacrée aux mémoires non volatiles, situe le dispositif étudié par rapport aux produits existants. Une compréhension des mécanismes de défaillance relatifs aux mémoires EEPROM nécessite une parfaite connaissance de leurs caractéristiques et de leur fonctionnement. Dans cette partie, le fonctionnement de la cellule EEPROM unitaire à deux niveaux de polysilicium est détaillé

ainsi que l'architecture et le fonctionnement de la matrice de cellules mémoires EEPROM. La deuxième partie du chapitre I débute par une description du test des circuits intégrés de manière générale. Nous verrons que le test d'un circuit se déroule obligatoirement en deux phases successives, dans une première phase des vecteurs de test ou « stimuli » sont appliqués sur les entrées du circuit et dans un deuxième temps les réponses observées sont comparées aux réponses attendues. De plus, le test doit être suffisamment exhaustif pour filtrer le maximum de défauts tout en restant le moins coûteux possible. Enfin, nous aborderons le test spécifique des mémoires non volatiles.

Le **deuxième chapitre** est consacré au diagnostic de défauts géométriques relatifs à la cellule mémoire EEPROM. Dans ce chapitre nous chercherons à mieux cerner l'influence de la variation des paramètres géométriques de la cellule EEPROM sur les valeurs des tensions de seuil. L'objectif étant la mise en place d'une méthodologie de diagnostic de défauts géométriques fiable qui sera validée sur silicium. Nous verrons que cette méthodologie de diagnostic se base sur un modèle électrique de cellule mémoire EEPROM développé à partir d'un modèle de transistor MM9 (MOS Model 9). Ce modèle électrique sera utilisé de manière à générer un modèle mathématique des tensions de seuil à partir d'une technique spécifique appelée plan d'expérience (Design Of Expriment). Le but final étant d'obtenir une équation polynomiale liant, d'une part, les trois tensions de seuil de la cellule et, d'autre part, des paramètres géométriques représentatifs de la cellule. La validation silicium de cette méthodologie de diagnostic innovante se fera à partir d'une cellule EEPROM fabriquée en technologie F6DP (technologie STMicroelectronics).
De manière à mieux comprendre l'élaboration du modèle de cellule mémoire EEPROM, un bref rappel sur le transistor MM9 ainsi que sur la modélisation du transistor MOS à grille flottante est donné au début du chapitre.

Dans le **troisième chapitre,** nous élargirons notre étude à la matrice de cellules mémoire EEPROM. Après une brève présentation des techniques d'analyses de défaillance classiques appliquées au plan mémoire EEPROM, nous aborderons la mise en place d'une méthode de diagnostic de défauts systématique qui a pour objectif d'établir une corrélation entre des signatures électriques représentatives de défauts simulés et les signatures électriques obtenues sur silicium après la phase de test.
Dans un premier temps, nous commencerons par définir une architecture d'étude composée d'une matrice de cellule mémoire élémentaire accompagnée de tous ses circuits périphériques. Ce circuit devra être à la fois représentatif du produit EEPROM et présenter des temps de simulation raisonnables. Dans un second temps, nous nous intéresserons tout particulièrement aux différents défauts à prendre en compte au niveau du circuit de simulation. Pour cela, un rappel sur la technologie étudiée (i.e. processus de fabrication) sera nécessaire de manière à mettre en évidence tous les défauts physiques pouvant potentiellement affecter le fonctionnement du circuit. Ensuite, une seconde étape sera consacrée à l'analyse et la modélisation de ces défauts. Enfin, une étude finale permettra la prise en compte au niveau du circuit de simulation de ces défauts.
Les résultats de simulation du circuit donneront un lien direct entre le défaut simulé et les signatures électriques correspondantes (tensions de seuil). L'objectif de cette analyse est de mettre en cause l'étape du processus de fabrication à l'origine de la défaillance uniquement à partir de la connaissance de signatures électriques.

En ce qui concerne les mémoires EEPROM, dont le principe de fonctionnement est basé sur un phénomène purement analogique, il apparaît clairement que la connaissance des trois tensions de seuil (dans l'état électrique écrit, effacé et vierge) de toutes les cellules d'un plan

mémoire est un paramètre clé dans le développement d'une méthodologie de diagnostic efficace. Or, les tests fonctionnels classiques appliqués aux mémoires EEPROM ne fournissent aucune information de nature analogique, alors qu'une telle information permettrait d'améliorer la phase de diagnostic. Dans les produits EEPROM standards, la tension de seuil de la cellule dans l'état écrit ne peut pas être directement mesurée, cependant nous verrons qu'une approche alternative va consister à extraire le courant de seuil I_{Tecrit} qui est l'image de la tension de seuil V_{Tecrit} sur la caractéristique I_d (V_{gc}) du transistor mémoire.

Le **quatrième chapitre** est consacré à la mise en place de moyens intégrés d'extraction des valeurs de seuil (courants et tensions) et de la séquence de test associée. Grâce à l'utilisation de ces moyens supplémentaires, on pourra extraire rapidement et avec précision les valeurs de seuil de toutes les cellules d'une mémoire EEPROM. A partir de ces données, nous développerons une méthodologie de diagnostic basée sur une cartographie bit analogique, pouvant être utilisée de concert avec la cartographie bit numérique classique (« bitmap »). Il en résulte une analyse plus efficace du comportement analogique de chaque cellule du plan mémoire. De plus, ces signatures électriques pourront servir d'entrée à un outil de diagnostic de défauts comme celui présenté au chapitre II.
La validation de ces structures d'extraction s'est déroulée en deux étapes. Tout d'abord, une étape de validation par simulation de la structure au sein d'un circuit mémoire élémentaire a été effectuée. Ensuite, une validation sur silicium a été réalisée à partir d'un véhicule de test constitué d'une matrice de cellules EEPROM 512Kbits fabriquée en technologie F8 CMOS 0.18 µm (technologie STMicroelectronics).

CHAPITRE I
Test des Mémoires Non Volatiles

I. Les Mémoires non volatiles

A. Les mémoires non volatiles à grille flottante de type EEPROM

1 Historique

Il existe plusieurs technologies de mémoires non volatiles. Lorsqu'il s'agit du stockage de charges, nous pouvons diviser les mémoires non volatiles en deux catégories : les mémoires à grille flottante et les mémoires à piégeage de charges [BRO98].

a. Mémoires non volatiles à grille flottante

Les travaux pour développer les points mémoires à grille flottante ont débuté dans les années 60. La volonté de réaliser des mémoires à semi-conducteurs avec de meilleures performances, était motivée par le succès et le rendement de la technologie MOS (Metal-Oxyde-Semiconductor). La nécessité d'adopter une solution permettant de sauvegarder l'information de façon permanente, même après coupure de l'alimentation, devenait évidente.
On a su résoudre ce problème en 1967, en proposant le concept de la grille flottante [KAH67] à travers les technologies MIMIS (Metal-Insulator-Metal-Insulator-Semiconductor) et MNOS (Metal-Nitrure-Oxide-Semiconductor) [WEG67]. La première mémoire à grille flottante a été validée et commercialisée en 1971 [BRO98], avec une capacité de 1Kbit, effaçable par rayons ultra-violets. Peu de temps après, une mémoire RAM (Random Acces Memory) de 1Kbit était introduite sur le marché. En 1979, la société Xicor réalise la première mémoire NOVRAM (Non Volatile RAM) [SAL99].
Dans les années 90, les mémoires PROM (Programmable Read Only Memory) ont été nettement améliorées. D'ailleurs, ce type de mémoire a constitué 10% du marché des mémoires à semi conducteur. Parallèlement au développement des technologies PROM, la première mémoire EEPROM (Electrically Erasable PROM), avec une capacité de 16Kbit a été commercialisée en 1983 en technologie MNOS.
Les mémoires Flash EEPROM ont ensuite pris un essor considérable durant la fin des années quatre vingt. Actuellement elles constituent la dernière génération de mémoires EEPROM et la troisième plus rapide croissance du marché des mémoires à semi-conducteurs, après les DRAM (Dynamic RAM) et les SRAM (Static RAM). Elles représentent une combinaison des bonnes propriétés des mémoires EEPROM et EPROM (Electrically Programmable ROM).
Aujourd'hui, les prédictions en terme de générations technologiques en microélectronique annoncent une limitation physique et technologique des composants silicium pour les années 2010-2015. Pour les transistors MOS, la limite de la longueur de grille est de l'ordre de 20 nm [PAL00]. Pour les mémoires non volatiles, l'épaisseur de l'oxyde à travers laquelle se fait l'injection de charges est limitée à 7 nm à cause de la rétention des cellules. Les perspectives envisagées sont alors orientées vers le développement des dispositifs mémoires à blocage de Coulomb [PAL00].

b. Mémoires non volatiles à piégeage de charges

Comme son nom l'indique, cette technique consiste à piéger des charges dans une couche diélectrique isolante. Plus précisément, la fonction non volatile de cette technologie est basée sur le stockage de charges dans des pièges localisés dans une couche de nitrure pour les

technologies MNOS (Metal-Nitride-Oxide-Semiconductor) et SNOS (Silicon-Nitride-Oxide-Semiconductor).
Ces charges sont injectées par effet tunnel à travers une fine épaisseur d'oxyde, de l'ordre de 1.5 à 3 nm. La quantité de charges injectées a pour effet de modifier la tension de seuil du transistor. Une cellule EAPROM (Electrically Alterable Programmable Read Only Memory) en technologie MNOS est présentée figure I.1. Dans cette structure, le transistor mémoire possède une grille conventionnelle en aluminium. Les charges sont injectées à travers la partie centrale en oxyde, entre le canal et la couche de nitrure SI_3N_4. C'est cette dernière qui joue le rôle de l'isolant à pièges de charges. Cette mémoire présente de nombreuses lacunes notamment en terme de rapidité et de densité d'intégration. De plus, elle nécessite l'application de 2 à 3 tensions différentes lors de son fonctionnement. En 1980, la technologie MNOS a été nettement améliorée avec la technologie SNOS. La fiabilité de cette technologie est basée sur l'utilisation de la méthode de dépôt LPCVD (Low-Pressure Chemical Vapor Deposition) pour le dépôt de polysilicium sur la couche de nitrure et la pré métallisation par un recuit d'hydrogène à haute température pour améliorer la qualité des interfaces nitrure-oxyde mince et oxyde mince-silicium [BRO98].

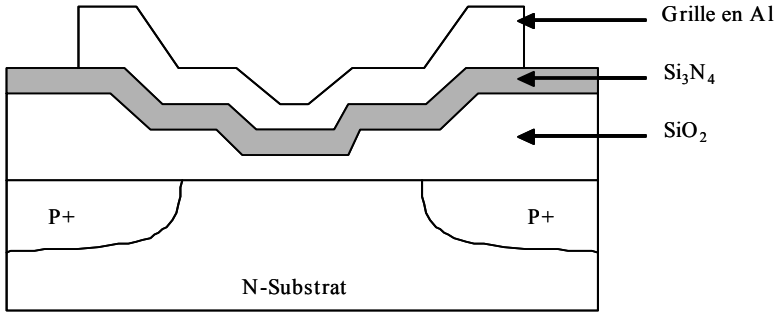


Figure I. 1 Structure en technologie MNOS.

Afin de réduire l'injection de charges de la grille vers le nitrure comme dans le cas de la technologie MNOS, la technologie SONOS (Silicon-Oxide-Nitride-Oxide-Semiconductor) a été proposée. Une couche d'oxyde (2 à 3 nm) est introduite entre le nitrure et la couche en polysilicium formant l'électrode de grille [CHE77]. Cette couche d'oxyde est obtenue soit par oxydation du nitrure (ce qui induit une réduction de l'épaisseur de nitrure), soit par dépôt. Un nouveau concept de la technologie SONOS a été mis en œuvre dans le but d'éviter la fuite de trous piégés dans l'oxyde de l'ONO vers l'électrode de grille tout en conservant une forte densité d'intégration ainsi que de basses tensions de programmation [SUZ83] [CHA87]. La nouvelle structure se compose d'une couche d'oxyde très mince, d'une fine épaisseur de nitrure (<10 nm) et d'un oxyde plus épais (>3 nm). Ceci permet la réduction de l'épaisseur totale de la structure et donc des tensions de programmation. Mais aussi une meilleure tenue en rétention de la cellule mémoire.

2 Technologie à grille flottante

a. Présentation

L'idée de base est de garder l'information même après coupure de l'alimentation électrique. Ceci est possible, lorsqu'on stocke des charges au niveau de la grille flottante d'un transistor mémoire, par exemple. Ces charges stockées ont pour effet de modifier la valeur de la tension de seuil du transistor. Par convention, cette tension de seuil peut être associée à deux valeurs logiques : un état logique 0 défini comme un état effacé et un état logique 1 défini comme un état écrit. La figure I.2 représente les caractéristiques électriques $I_d(V_{gc})$ du transistor mémoire

dans l'état écrit et effacé. La détection de l'état logique du transistor mémoire est obtenue en appliquant une tension de lecture de grille comprise entre les tensions de seuil $V_{Tefface}$ et V_{Tecrit}. Ainsi, pour ces conditions de lecture, le transistor est soit dans un état passant, soit dans un état bloqué. Le maintien des charges stockées, même après coupure de l'alimentation électrique garantit la non volatilité de la mémoire.

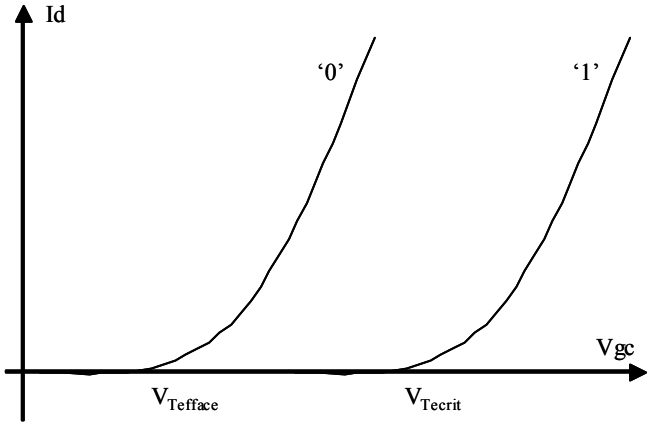


Figure I. 2 Tensions de seuil et états logiques.

b. Technologie

Le point commun entre les mémoires EPROM, EEPROM et Flash EEPROM est l'utilisation du principe du transistor de mémorisation à grille flottante [CAP99]. Le principe de fonctionnement de ces mémoires sera détaillé dans le paragraphe suivant.
Le fonctionnement du transistor de mémorisation à grille flottante consiste à stocker des charges électriques dans un matériau conducteur ou semi-conducteur en faisant passer ces charges à travers un diélectrique (S_iO_2). La différence entre les technologies à grille flottante est liée aux mécanismes d'injection de charges utilisés durant les phases d'écriture et d'effacement.
Lorsque le mécanisme d'injection par porteurs chauds est utilisé en phase de programmation (écriture), la cellule est considérée comme étant de technologie SIMOS (Stacked gate Injection MOS). Cette technologie est utilisée principalement dans le cas des mémoires de type EPROM.
Dans le cas où le mécanisme d'injection utilisé est de type Fowler-Nordheim à travers un oxyde mince (7 à 10 nm), la technologie est appelée FLOTOX (FLOating gate Thin OXide). Cette technologie est souvent utilisée dans les applications de type EEPROM, Flash EEPROM et NOVRAM.
Une troisième technologie est appelée TPFG («Textured Poly Floating Gate »). Le mécanisme utilisé est de type Fowler-Nordheim, comme la technologie FLOTOX, mais à travers une couche en poly-oxyde [KLE79] [LAN80].
Le mécanisme d'injection Fowler-Nordheim a été utilisé pour la première fois, en mode écriture et effacement, dans une cellule mémoire non-volatile de type RAM [HAR78]. Cette technologie a été à l'origine de l'apparition de la mémoire EEPROM. La structure FLOTOX est présentée figure I.3. La première grille, entourée de diélectrique, est appelée grille flottante [KAH67], la deuxième est appelée grille de contrôle. On utilise généralement une structure oxyde-nitrure-oxyde (ONO) pour le deuxième diélectrique situé entre la grille flottante et la grille de contrôle. Une couche d'oxyde mince sépare le drain de la grille flottante. Sous l'effet d'un champ électrique intense de l'ordre de 10 MV/cm, les électrons passent par effet tunnel à travers l'oxyde mince, entre le drain et la grille flottante. La tension de seuil du transistor à

grille flottante s'en trouve alors modifiée. Cette tension de seuil augmente si on applique un signal haute tension (typiquement de l'ordre de 12V) sur l'électrode de grille de contrôle. Elle diminue si on applique un signal haute tension sur l'électrode de drain.

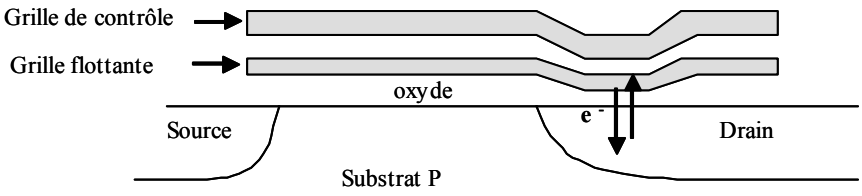


Figure I. 3 Structure mémoire en technologie FLOTOX.

3 Les différentes familles de mémoires non volatiles

a. La mémoire ROM

Ce type de mémoire est codé soit au cours du procédé de fabrication (activation ou non de transistors par masquage), soit par l'utilisateur avec des structures à base de fusibles [CRR96]. Un point mémoire ROM peut être codé suivant deux états : passant ou bloqué. Un point mémoire ROM passant est un simple transistor NMOS. Un point mémoire ROM bloqué est un transistor NMOS qui ne pourra jamais être passant quelles que soient les polarisations de grille ou de drain appliquées.
Il existe différentes méthodes de programmation des mémoires ROM bloquées :

- codage lors de l'implant drain extension, N-LDD (Lightly Doped Drain) des transistors NMOS. Dans ce cas le masque avant l'implant de phosphore est tel que seule la source soit implantée (figure I.4.a),
- codage lors de l'implant P-LDD relatif aux transistors PMOS. Ici, les sources et drains ont été implantés lors d'une étape antérieure par l'implant de phosphore N-LDD. Lors de l'implant de bore P-LDD, la source de la ROM n'est pas protégée par un masque de résine. L'implant de bore vient inhiber la source des ROM bloquées (figure I.4.b),
- codage lors de la croissance de l'oxyde de champ. Ce type de codage se fait au début du processus de fabrication. L'oxyde de champ vient croître sur la zone de drain des ROM bloquées (figure I.4.c).

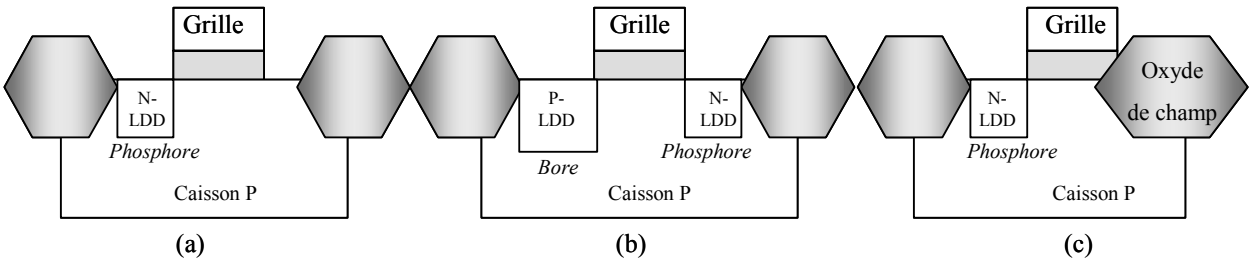


Figure I. 4 Codage des mémoires ROM.

b. La mémoire EPROM

La cellule EPROM est un transistor NMOS avec une grille flottante isolée entre le substrat et la grille de contrôle. La grille flottante étant complètement ancrée dans l'oxyde comme le montre la figure I.5.

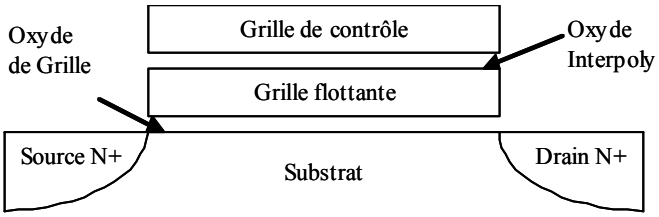


Figure I. 5 Vue en coupe d'une mémoire EPROM.

Une cellule EPROM est écrite électriquement, mais effaçable par rayons ultraviolets [BRO98], la phase d'écriture peut durer de 100 µs à 1 ms, la phase d'effacement demande 20 minutes d'exposition aux ultraviolets. L'écriture d'une telle cellule se fait par injection de porteurs chauds.
Le principe de fonctionnement d'une cellule EPROM est basé sur l'accumulation de charges dans la grille flottante. Ces charges ainsi piégées vont engendrer un décalage de la tension de seuil du transistor mémoire. L'opération d'effacement par rayons ultraviolets (UV) reste lourde à mettre en œuvre puisqu'elle suppose un démontage du boîtier de son support et un passage sous rayons ultraviolets. Les mémoires EPROM utilisent des boîtiers coûteux à fenêtre à quartz.

c. La mémoire EEPROM

La figure I.6 montre que ce type de mémoire se compose de deux transistors. Un transistor de sélection qui permet d'adresser le point mémoire et de sélectionner le mode d'utilisation (programmation ou lecture) [YAR92]. Un transistor MOS à grille flottante appelé transistor d'état ou transistor mémoire qui permet de stocker l'information sous forme de charges électriques.

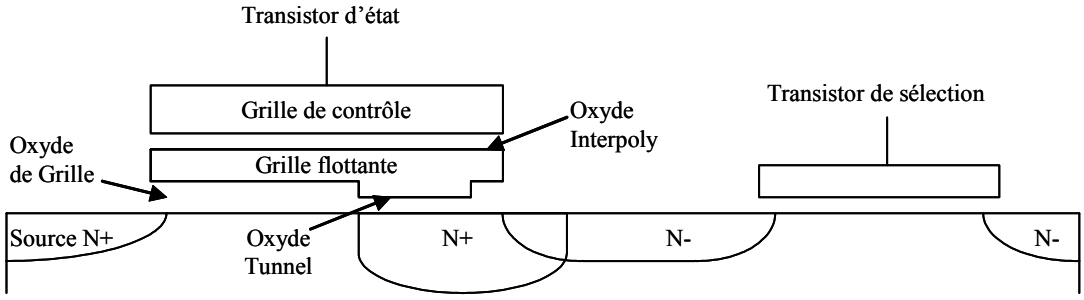


Figure I. 6 Vue en coupe d'une mémoire EEPROM.

La sélection du point mémoire se fait lorsque le transistor de sélection devient passant. Dans ce cas, la cellule peut être utilisée soit en mode lecture, soit en mode de programmation.
L'écriture ou l'effacement de la cellule EEPROM se fait par injection de charges à travers l'oxyde tunnel. C'est une injection par courant tunnel Fowler Nordheim (FN) qui a lieu à fort champ et à travers une épaisseur d'oxyde inférieure à 10 nm.
On distingue trois états différents de la cellule : vierge, effacé et programmé. La cellule ne peut se trouver dans l'état vierge qu'après effacement sous UV. Lors de son utilisation, la cellule EEPROM est soit dans un état vierge soit dans un état programmé. Nous développerons plus en détail le fonctionnement de la mémoire EEPROM dans les paragraphes suivants.

d. La mémoire Flash EEPROM

La cellule mémoire Flash EEPROM (appelée plus simplement mémoire Flash) se compose d'un unique transistor MOS à grille flottante comme le montre la figure I.7. L'idée étant d'avoir un seul transistor mémoire qui offre à la fois une rapidité de programmation, une haute densité d'intégration et un effacement électrique similaire à la cellule EEPROM.

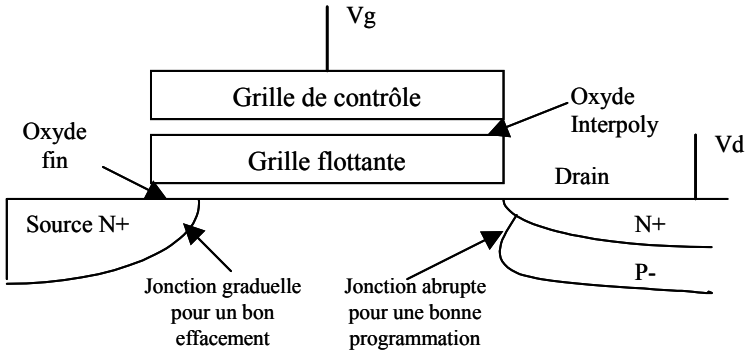
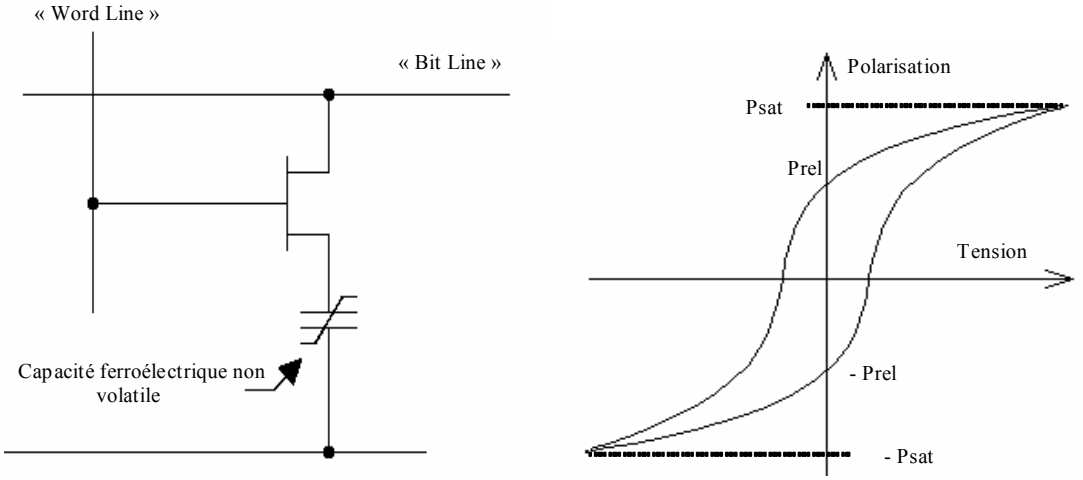


Figure I. 7 Vue en coupe d'une mémoire Flash EEPROM.

La programmation de la cellule s'effectue par porteurs chauds [CAR89]. Cette opération a pour but de stocker une quantité fixée de charges dans la grille flottante de la cellule. Les électrons sont accélérés le long du canal par une forte tension de drain (V_d = 5V typiquement) et acquièrent une énergie importante. Les chocs avec le réseau cristallin au voisinage du drain créent de nouvelles paires électrons-trous. C'est le phénomène d'ionisation par impact.
Durant cette phase, les électrons sont attirés vers la grille de contrôle fortement positive (tension de l'ordre de 10V). L'énergie importante de ces électrons au voisinage du drain rend possible le franchissement de la barrière de potentiel S_i/S_iO_2 et leur permet d'atteindre la grille flottante. Ce type d'écriture est très rapide (quelques microsecondes) mais requiert beaucoup d'énergie (300 µA de courant de drain pendant l'écriture, à V_d=5 V, pendant 5 µs). Il reste de toute manière bien plus rapide que l'injection FN (quelques millisecondes) car des considérations en terme de fiabilité imposent de rester à des champs électriques FN modérés.
L'écriture par porteurs chauds de la cellule nécessite une jonction de drain abrupte afin de générer des électrons de forte énergie et de favoriser leur injection dans la grille flottante. Elle est réputée plus fiable que l'injection FN : le champ électrique imposé dans l'oxyde est plus faible [EIT96].
L'application d'une tension fortement positive sur la source (en maintenant la grille de contrôle à un potentiel nul) permet aux électrons stockés dans la grille flottante de passer à la source par conduction FN localisée [HUF80]. Durant cette opération, le drain est généralement laissé flottant de manière à limiter le courant de fuite entre drain et source.
Ce type d'effacement nécessite d'une part un large recouvrement entre la jonction de source et la grille flottante (double diffusion) afin de limiter la génération de « porteurs chauds » dans la jonction source-substrat.
La lecture du point mémoire se fait par la détection du courant de lecture pour un état de polarisation donné (V_d=1V, V_g=5V). Ce courant de lecture dépend de la quantité de charges stockées dans la grille flottante (il atteint 50 µA). Les fortes capacités de drain (liées aux capacités des jonctions drain-substrat) limitent la vitesse de lecture. Au niveau circuit, l'opération de lecture nécessite généralement une opération de précharge. La vitesse de lecture suivant les produits varie entre 50 ns et 150 ns.

e. Mémoires émergentes

Un nouveau type de mémoires non volatiles est promis à un bel avenir : les mémoires ferroélectriques. Elles présentent des caractéristiques intéressantes mais des obstacles techniques et un manque d'investissement a retardé la révolution qu'elles promettent.
Ce type de mémoire s'obtient par déposition d'une couche de matériaux ferroélectriques cristallins entre deux électrodes de manière à former une capacité. L'effet ferroélectrique peut être défini comme la capacité à maintenir un état de polarisation en l'absence de champ électrique [TOY80]. Le point mémoire FRAM est le plus souvent représenté par une capacité ferroélectrique accompagnée d'un transistor d'accès (figure I.8a).
Le mécanisme de mémorisation de ces cellules est différent de celui utilisé dans les mémoires non volatiles. Le cycle d'hystérésis du point mémoire FRAM montre l'évolution du niveau de polarisation de la capacité ferroélectrique en fonction du champ électrique appliqué à ses bornes. Ce cycle d'hystérésis est représenté figure I.8b. L'augmentation du champ électrique aux bornes de la capacité ferroélectrique augmente son niveau de polarisation jusqu'à saturation à la valeur P_{sat}. Un phénomène similaire se produit lorsque le champ électrique est inversé (saturation à la valeur $-P_{sat}$). On peut donc mettre en évidence deux états stables de polarisation, au niveau des valeurs de relaxation P_{rel} et $-P_{rel}$. Ces états ne nécessitent aucun champ électrique ni courant pour être maintenus [TAK97].
La lecture du point mémoire consiste à mesurer le courant nécessaire pour passer la cellule de son point de relaxation (P_{rel} ou $-P_{rel}$) à l'état de polarisation positif P_{sat}.
Il faut un faible courant pour passer de l'état P_{rel} à P_{sat}, l'information lue est un 0 logique. Par contre un fort courant est nécessaire pour passer de l'état de relaxation $-P_{rel}$ à l'état P_{sat}, l'information lue est un 1 logique. Notons qu'après avoir lu un 1 logique, il faut réécrire le point mémoire pour qu'il retourne à son état de relaxation négatif $-P_{rel}$.
Le temps de programmation d'une mémoire FRAM est de 100 ns, soit dix fois moins que pour une mémoire Flash EEPROM. De plus, ces mémoires consomment peu d'énergie.

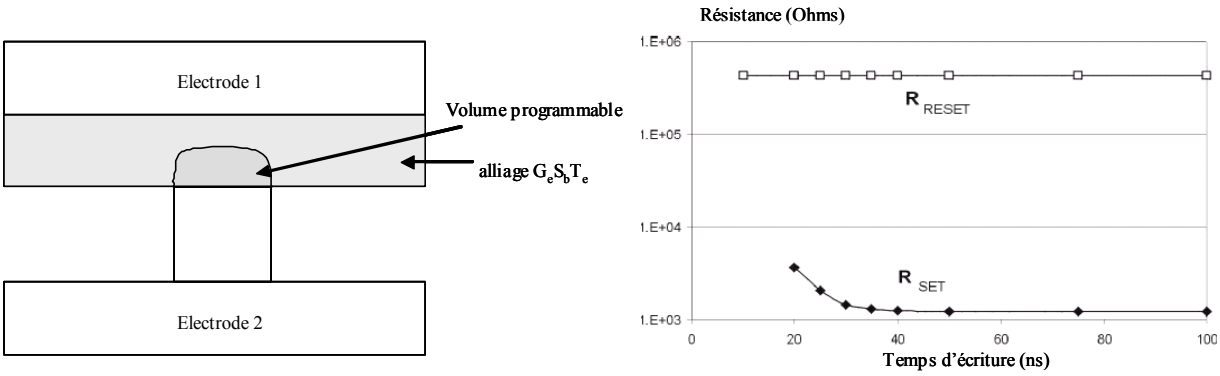


(a) Le point mémoire FRAM (b) Cycle d'hystérésis FRAM

Figure I. 8 Structure et fonctionnement des FRAM.

Cependant, les matériaux ferroélectriques constituent aussi la faiblesse des FRAM. En effet, les programmations successives entraînent une usure des matériaux, et l'endurance de ces mémoires s'en trouve affectée. De plus, un deuxième problème rencontré est l'incompatibilité entre le procédé CMOS et le matériau ferroélectrique. Résoudre ces problèmes d'endurance et d'intégration constitue donc un enjeu majeur dans le développement des FRAM.

Une technologie récente, la technologie OUM (Ovonic Unified Memory), est censée présenter des avantages par rapport aux technologies de mémoires non volatiles existantes, à savoir un coût réduit, des temps d'accès en écriture et lecture plus rapide, une endurance et une évolutivité améliorées. La technologie OUM est une version électrique du processus de changement de phase réversible de l'état cristallin à l'état amorphe. La technologie OUM fut annoncée dès la fin des années 60 mais ne fut pas commercialisée suite à certains problèmes techniques dans le domaine des matériaux.
Les mémoires OUM utilisent les propriétés de changement de phase d'un alliage $G_eS_bT_e$ [WIC99]. Ce changement de phase est obtenu par élévation de température d'un petit volume de matériau traversé par un courant électrique. Il en résulte un changement considérable de la résistivité du matériau.
La phase amorphe, de forte résistivité est définie comme l'état « RESET » et l'état poly cristallin, caractérisé par une faible résistivité est défini comme l'état « SET » (figure I.9).



(a) vue en coupe d'une mémoire FRAM

(b) Etats programmés « SET » et « RESET »

Figure I. 9 Structure et fonctionnement des mémoires OUM.

Pour avoir de bonnes performances en lecture et écriture, certains fabricants de semi-conducteurs proposent des mémoires non volatiles à base de cellules RAM. La plus complexe de ces mémoires est la NOVRAM (Non Volatile RAM) dans laquelle une mémoire EEPROM est couplée à une RAM dynamique ou statique.
D'autres NOVRAM sont constituées d'une mémoire RAM et d'une alimentation de sauvegarde intégrée dans le même boîtier.

4 Caractéristiques des MNV de type EEPROM

Au delà de leurs caractéristiques intrinsèques, les mémoires EEPROM possèdent d'importantes caractéristiques fonctionnelles. Ces dernières permettent d'évaluer les performances de la cellule EEPROM. Ces caractéristiques se divisent en trois catégories principales : les caractéristiques transitoires, en endurance et en rétention.

a. Caractéristiques transitoires

Le principe de fonctionnement de ce type de mémoire repose sur la possibilité d'injecter des charges dans la grille flottante et aussi de décharger cette dernière, entraînant un décalage de la tension de seuil de la cellule.
Les caractéristiques transitoires décrivent l'évolution de la tension de seuil en fonction de la durée et de la forme des signaux de programmation. Une parfaite connaissance de ces

caractéristiques permet de déterminer et d'optimiser les tensions ainsi que les temps de programmation. Mais aussi, d'améliorer la rétention des cellules en limitant l'intensité du champ électrique aux bornes de l'oxyde tunnel [CAN00].

b. Caractéristiques en endurance

Les mémoires non volatiles peuvent être reprogrammées. Mais, contrairement aux mémoires RAM, chaque programmation introduit « une usure » permanente au niveau du point mémoire. Cela implique une limitation du nombre total de cycles de programmation (10^5 cycles pour les produits EEPROM actuels). Le nombre de cycles supportés par la mémoire avant l'apparition de la première défaillance porte le nom d'endurance.
Le test d'endurance consiste à effectuer plusieurs cycles d'effacement-écriture sur une cellule élémentaire et prélever la tension de seuil (en écriture et en effacement) correspondant à chaque nombre de cycles. La différence entre la tension de seuil en écriture et celle en effacement est appelée fenêtre de programmation.
L'endurance d'une mémoire EEPROM est principalement limitée par la tension de claquage de l'oxyde tunnel. Le claquage d'oxyde serait provoqué par des points faibles (défauts) situés dans l'oxyde. Ces défauts sont généralement activés par de forts champs électriques et engendrent des fuites de charges.
Un autre phénomène est lié au piégeage de charge dans l'oxyde durant les phases de programmation. Il a pour effet de modifier le champ aux bornes de l'oxyde et dans le même temps la quantité de charges injectées. Cela peut induire un décalage insuffisant des tensions de seuil et perturber la détection de l'état de la cellule durant la phase de lecture [MIE87].

c. Caractéristiques en rétention

Le but du test en rétention est d'estimer la capacité de la cellule mémoire à sauvegarder l'information. Dans les conditions d'utilisation de la cellule, c'est-à-dire à température ambiante, la cellule ne doit pas perdre plus de 10% de la charge stockée en 10 ans. Afin d'activer et d'accélérer la perte de charge, le test en rétention se fait généralement à haute température (200°C, 250°C, 300°C). On effectue par la suite une mesure de la dérive de la tension de seuil à différentes températures. L'extrapolation de cette dérive à température ambiante, suivant la loi d'Arrhenus, permet d'évaluer la durée de vie de la cellule mémoire [SAL99].

B. Architecture et fonctionnement des mémoires EEPROM

1 La cellule mémoire EEPROM : principe de fonctionnement

a. L'injection Fowler-Nordheim

Les opérations d'écriture et d'effacement de la cellule EEPROM sont basées sur le mécanisme d'injection tunnel Fowler-Nordheim (FN) [BRO98] [LENZ69]. Le concept de l'effet tunnel à travers une barrière de potentiel s'applique aux structures présentant un oxyde mince [EVT98].
La probabilité qu'un électron traverse la barrière de potentiel par effet tunnel dépend de la distribution des états occupés dans le matériau d'injection, de la forme, la taille et la largeur de la barrière. En utilisant le modèle de l'électron libre pour le métal et l'approximation de

Wentzel-Kramers-Brillouin (WKB) pour la probabilité de traverser la barrière de potentiel, on obtient l'expression de la densité de courant tunnel suivante :

$$J_{FN} = \frac{q^3 E^2}{16\pi^2 h^2 \Phi_B} . exp\left[\frac{(-4.(2m_{ox}{}^*)^{1/2} . \Phi^{3/2}}{3\hbar qE} \right] \text{(I. 1)}$$

où Φ_B la hauteur de barrière, $m_{ox}{}^*$ la masse effective, ħ la constante de Planck modifiée, q la charge de l'électron, et E le champ électrique à travers l'oxyde.
Le champ électrique est égal à la tension appliquée divisée par l'épaisseur d'oxyde. Une réduction de l'épaisseur d'oxyde sans réduction proportionnelle de la tension appliquée produit une augmentation rapide du courant tunnel. Avec de l'oxyde relativement épais (20 à 30 nm) on doit appliquer une tension élevée (20 à 30 V) pour avoir un courant tunnel. Avec des oxydes minces, le même courant peut être obtenu en appliquant une tension beaucoup plus basse. Une épaisseur optimale d'environ 80 Å est choisie dans les dispositifs actuels qui utilisent le phénomène d'injection par effet tunnel. Cette épaisseur est un compromis entre les contraintes de performance (vitesse de programmation, consommation d'énergie, ...), qui exigeraient des oxydes minces, et les contraintes de fiabilité, qui exigerait des oxydes épais.
L'utilisation du mécanisme d'injection par effet tunnel est justifié par les raisons suivantes : il permet d'obtenir des temps de programmation inférieurs à 1 ms et une rétention supérieure à 10 ans, ce qui est fondamental pour toutes les technologies non volatiles. Ce phénomène est purement électrique et le niveau des courants induits est très faible, ce qui permet la génération des tensions d'alimentation requises pour toutes les phases de fonctionnement des EEPROM.
Cependant, la dépendance exponentielle du courant tunnel par rapport au champ électrique pose quelques problèmes critiques comme le contrôle du processus de fabrication. En effet, une faible variation de l'épaisseur d'oxyde parmi les cellules d'une matrice mémoire provoque une grande variation de courant de lecture. De ce fait, la distribution des tensions de seuil d'une matrice est très dispersée. Un contrôle du processus de fabrication de très bonne qualité est donc exigé.
La fiabilité des oxydes minces est devenue une des préoccupations principales des fabricants de circuits intégrés. La durée de vie des oxydes a un impact direct sur les propriétés de rétention de charges des mémoires non volatiles. Dans ce contexte, une meilleure connaissance des défauts dans l'oxyde et de leurs mécanismes de création est devenue nécessaire [MAN99] [SAM95].
Les oxydes de mauvaise qualité sont riches en pièges et défauts d'interface. La conduction par effet tunnel est de ce fait améliorée puisque la taille équivalente de la barrière de potentiel vue par des électrons est réduite. Ainsi, le passage par effet tunnel nécessite un champ aux bornes de l'oxyde fin inférieur à 10 MV/cm.
Un comportement en programmation stable et homogène sur tout le plan mémoire nécessite une densité de défauts extrêmement faible. Des écritures ou des effacements fréquents induisent une augmentation de charge emprisonnée dans l'oxyde. Ceci affecte la taille de la barrière de potentiel de l'oxyde tunnel, qui diminue dans le cas d'accumulations positives et augmente dans le cas d'accumulations négatives à l'interface entre l'oxyde et le semi-conducteur. Cela se traduit par une variation des courants tunnels [YAM96].

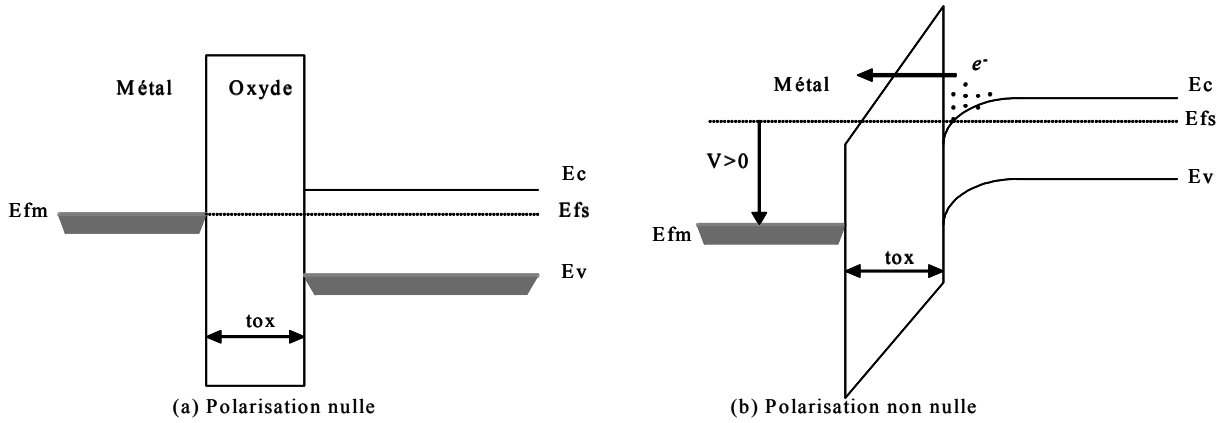


Figure I. 10 Diagramme des bandes d'énergie d'une structure MOS.

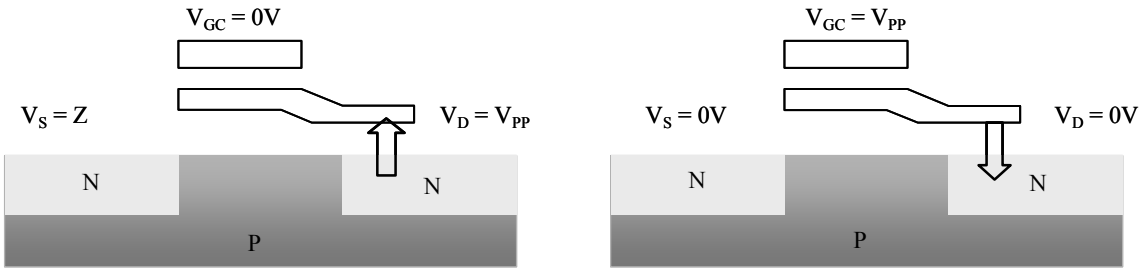
Un diagramme des bandes d'énergie de la structure MOS à polarisation nulle et non nulle est présenté figure I.10. La barrière de potentiel est réduite sous l'effet d'une forte polarisation positive, donnant naissance à un courant tunnel à travers l'oxyde. Suivant le sens du champ électrique induit par la polarisation, la grille métallique se charge négativement ou positivement.
Bien que la forme simple et classique de la densité de courant tunnel soit en bon accord avec les données expérimentales, beaucoup de cellules sont mal évaluées à cause de la dépendance en température du phénomène, des effets quantiques à l'interface du silicium, de l'influence de l'effet de bande entre le substrat et le drain à l'interface S_i/S_iO_2, de la chute de tension dans le silicium, et du fait que les statistiques correctes pour des électrons ne sont pas de Maxwell mais plutôt de Fermi-Dirac… Ces paramètres ont d'ailleurs une grande importance pour la modélisation et la simulation de la cellule mémoire EEPROM.
Il est possible de réécrire l'expression de la densité de courant sous une formule plus simple mais identique :

$$J = A.E^2.\exp\left[\frac{-B}{E}\right] \text{(I. 2)}$$

A et B étant fonction du champ électrique et incluant les effets quantiques. Cette approche est tout à fait satisfaisante dans beaucoup de cas mais mène à différentes valeurs de A et B selon l'électrode d'injection et la polarisation de la cellule.

b. Ecriture de la cellule EEPROM

La phase d'écriture, présentée figure I.11a, consiste à appliquer une impulsion haute tension V_{pp} sur le drain du transistor mémoire EEPROM, la grille de contrôle ainsi que le substrat se trouvant à la masse. La grille du transistor de sélection (non représenté) est polarisée à un potentiel d'amplitude supérieure à celle de l'impulsion appliquée sur le drain, ce qui garantit l'état passant du transistor de sélection. La source peut être soit connectée au drain de la cellule, soit laissée à un potentiel flottant. La première solution est possible lorsqu'il s'agit d'une cellule élémentaire. Cependant, lorsqu'il s'agit d'une matrice mémoire, la deuxième solution est adaptée pour la phase d'écriture.
Dans les deux cas, les électrons sont évacués de la grille flottante vers le drain, entraînant une diminution de la tension de seuil du transistor mémoire. La grille flottante se retrouve ainsi chargée positivement.



(a) Phase d'écriture (b) Phase d'effacement

Figure I. 11 Programmation de la cellule EEPROM.

c. Effacement de la cellule EEPROM

Pendant la phase d'effacement, une impulsion haute tension est appliquée sur la grille de contrôle du transistor mémoire. Le drain, la source ainsi que le substrat étant reliés à la masse. La phase d'effacement correspond à une injection d'électrons du drain vers la grille flottante comme le montre la figure I.11.b, entraînant une augmentation de la tension de seuil du transistor mémoire.

d. Lecture de la cellule EEPROM

Une cellule est écrite ou effacée lorsque sa tension de seuil est respectivement plus faible ou plus élevée que celle de la cellule mémoire vierge. La figure I.12 représente les trois états électriques d'une cellule mémoire EEPROM.
Pour lire l'état d'un point mémoire, il suffit d'appliquer une tension sur la grille de contrôle V_{gc} située entre la tension de seuil d'une cellule programmée et la tension de seuil d'une cellule écrite. Il en résulte l'existence ou l'absence de courant suivant que la grille flottante est chargée négativement ou positivement. Au niveau circuit, ce sera l'amplificateur de lecture qui détectera le courant fourni par la cellule adressée.

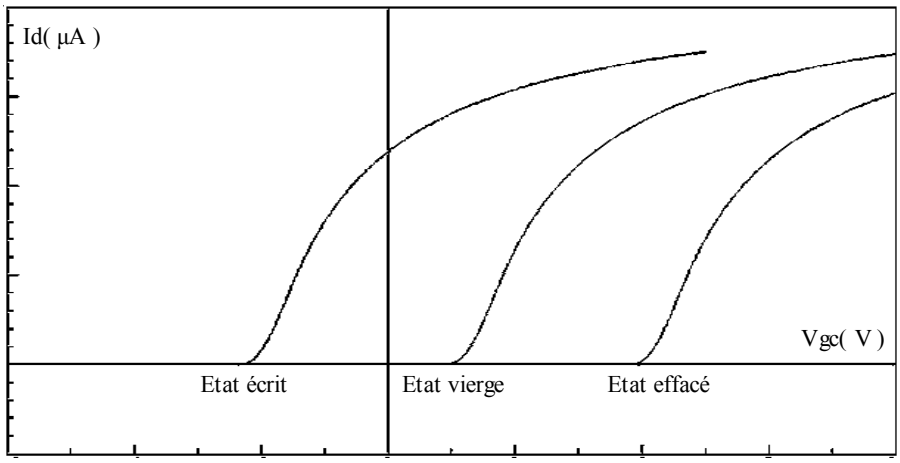


Figure I. 12 Les trois états électriques de la cellule EEPROM.

e. Fiabilité

Un composant est fiable lorsqu'il est capable d'assurer son bon fonctionnement pendant la durée de son utilisation. La fiabilité dépend essentiellement des caractéristiques physiques (type de matériaux) et de la technologie (technique de fabrication) du dispositif utilisé. Les technologies à grille flottante sont soumises à des champs électriques relativement forts,

pouvant dégrader la zone d'injection tunnel. Ceci constitue le problème majeur de la fiabilité des mémoires non volatiles de type EEPROM.
Une cellule EEPROM est fiable lorsqu'elle est capable de conserver l'information pendant plusieurs années (10 ans). Mais elle doit aussi assurer un bon fonctionnement après plusieurs cycles d'écriture et d'effacement, ce qui définit l'endurance des mémoires EEPROM.

2 Le plan mémoire EEPROM

a. Architecture de type NOR

Le plan mémoire EEPROM est une matrice de lignes (« word line ») et colonnes (« bit line »). A chaque intersection correspond un point mémoire. Dans la plupart des cas, l'architecture NOR, présentée figure I.13, est utilisée.
Pendant les opérations de lecture, la cellule à lire est adressée en positionnant sa ligne à une tension positive alors que les autres lignes de la matrice sont connectées à la masse. Cependant pour les mémoires Flash, à point mémoire de type ETOX, il faut que toutes les cellules non sélectionnées aient une tension de seuil équivalente suffisamment positive (ce qui entraîne un courant nul sur la « bit line » lorsque les grilles des cellules non sélectionnées se trouvent à un potentiel nul). Dans le cas contraire, des courants de fuite peuvent apparaître, ce qui perturbe énormément le signal de lecture sur la « bit line ». L'architecture NOR est donc particulièrement sensible à l'hyper effacement des mémoires Flash.

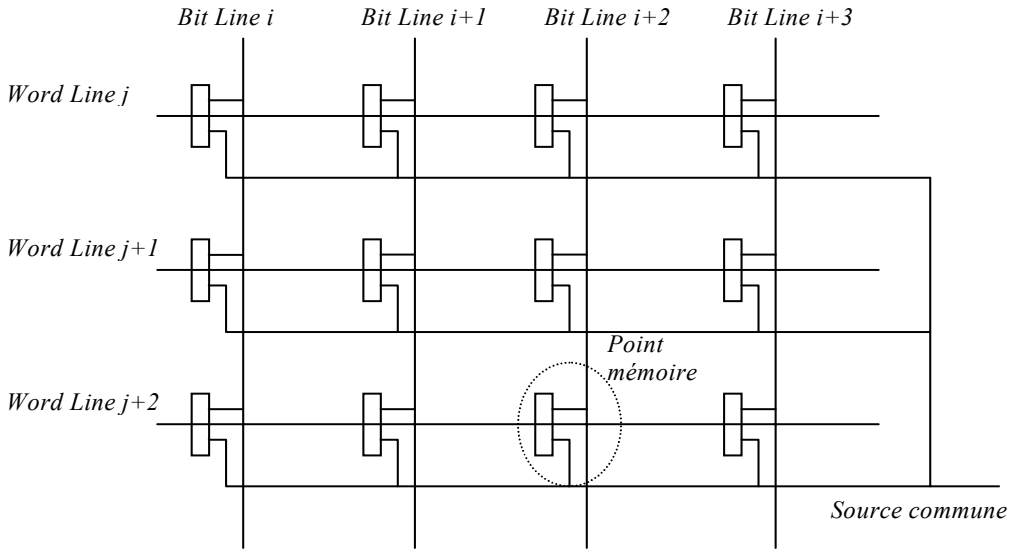


Figure I. 13 Architecture de type NOR.

Le principal inconvénient des architectures NOR est leur relative faible densité d'intégration. En effet, tous les points mémoires ont leur drain connecté à la « bit line » ainsi que leur électrode de source à une ligne commune. Ceci génère un nombre important de contacts.
Dans les mémoires EEPROM de type NOR, toutes les cellules mémoires sont reliées aux « bit lines » à travers des transistors de sélection comme le montre la figure I.14. De plus, un transistor de sélection supplémentaire est ajouté pour chaque mot mémoire de manière à pouvoir accéder en effacement à un seul mot par page contrairement aux mémoires Flash EEPROM.

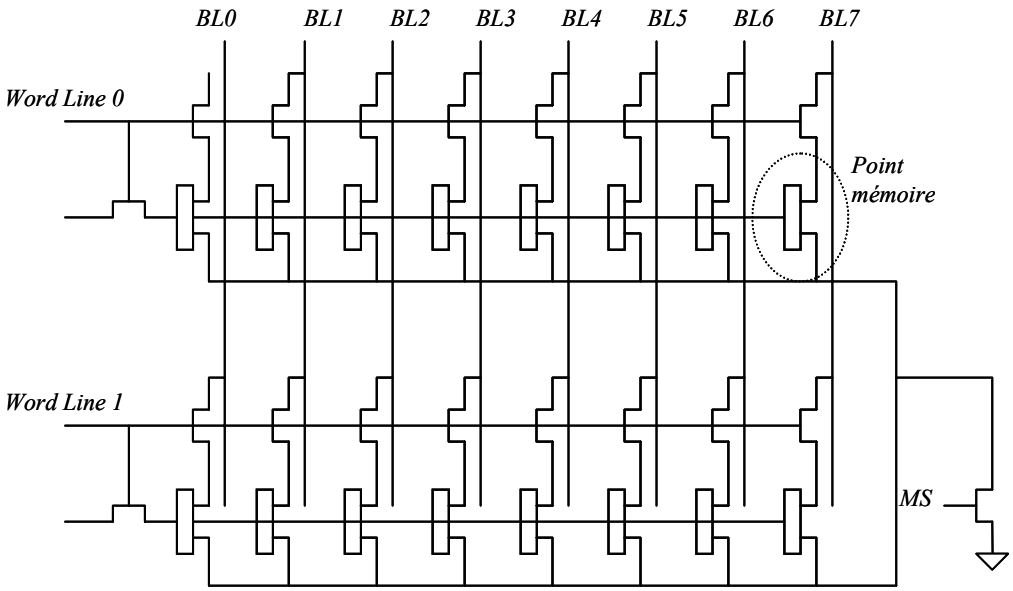


Figure I. 14 Architecture NOR des mémoires EEPROM.

Dans les mémoires EEPROM de type FLOTOX, 8, 16 ou encore 32 bits adjacents forment un mot. La manière dont sont connectés ces mots (toutes les sources des transistors mémoire sont connectées entre elles) rend impossible une programmation sélective de quelques bits à l'intérieur d'un mot. Par conséquent, la programmation d'un bit au sein de cette architecture s'effectue en deux étapes.
Tout d'abord, les tensions de seuil de tous les transistors du mot sélectionné sont simultanément portées à des valeurs hautes. Ensuite certains bits, préalablement sélectionnés, sont écrits en diminuant la valeur de leur tension de seuil.
L'opération qui permet une modification simultanée de l'état électrique de tous les bits d'un mot est appelée effacement. L'opération qui permet un accès sélectif à certains bits est une opération d'écriture. L'opération d'effacement entraîne donc une augmentation de la tension de seuil d'un mot mémoire complet vers une valeur positive, alors qu'une opération d'écriture entraîne une diminution de la tension de seuil de certains bits spécifiques d'un mot vers une valeur inférieure à zéro. Cette convention, adoptée pour une grande majorité de produits EEPROM, sera utilisée dans cet ouvrage.

b. Architecture de type NAND

Une meilleure densité d'intégration est obtenue grâce à une architecture de type NAND (figure I.15), uniquement valable pour les mémoires flash EEPROM. Dans ce cas, les « bit lines » sont composées de points mémoires connectés en série. Chaque colonne comprend deux transistors de sélection, le premier commandé par le signal SL (« Select Line »), permet de sélectionner la colonne adressée. Le second commandé par le signal GS (« Ground Select ») permet de relier les colonnes à la masse. Pour accéder à n'importe quelle cellule de la ligne, il faut non seulement activer la « word line » de la cellule sélectionnée, mais aussi toutes les « word lines » commandant les autres cellules de la ligne.

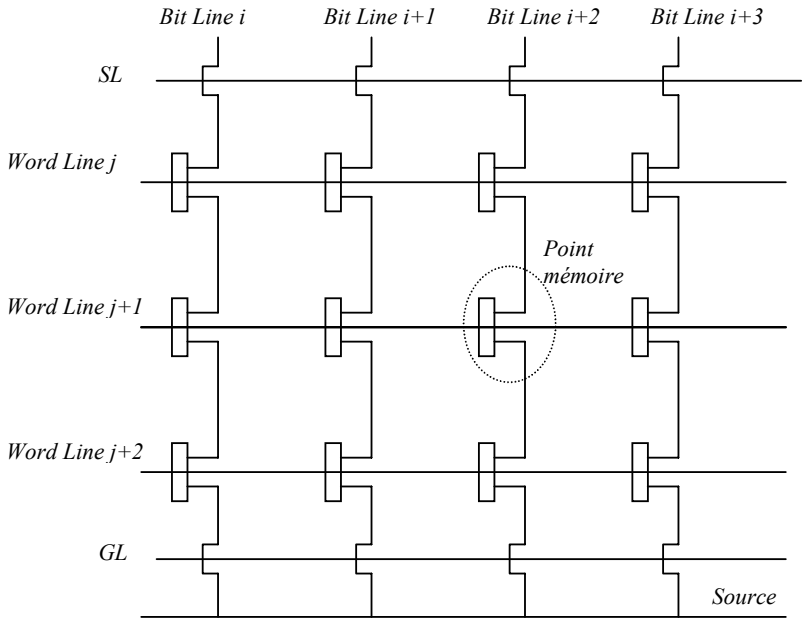


Figure I. 15 Architecture de type NAND.

Lors d'une opération de lecture, on applique sur la grille du point mémoire sélectionné une tension de lecture assez faible (proche de zéro), alors qu'une tension de grille supérieure à la tension de seuil maximale de la technologie considérée est appliquée aux autres points mémoire. Ainsi la cellule sélectionnée impose le courant de « bit line » à lire.

3 Fonctionnement des mémoires EEPROM

La structure des mémoires EEPROM obéit à deux principes généraux :

- l'exploitation du plan mémoire qui nécessite l'élaboration d'une électronique analogique chargée de le mettre en œuvre (en tenant compte des contraintes imposées par la technologie) et de convertir l'état des cellules mémoires en grandeurs binaires.
- le traitement de l'information binaire, réalisé par une logique de contrôle chargée de la gestion de la communication entre le plan mémoire, les circuits périphériques et les entrées-sorties du composant.

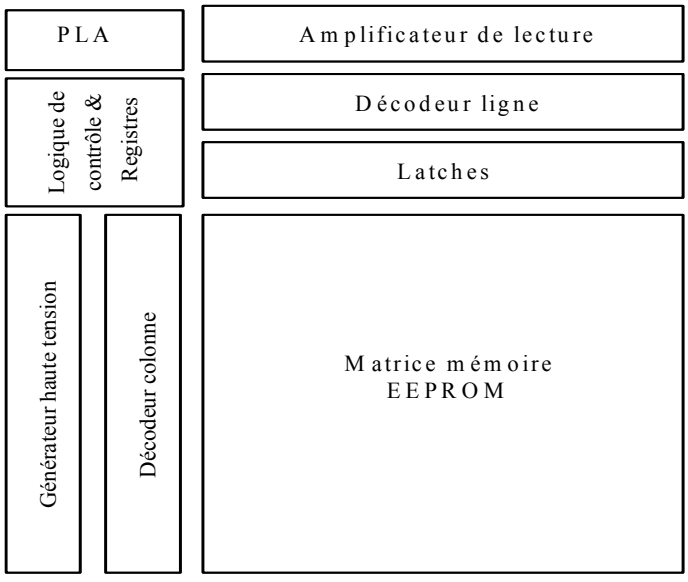


Figure I. 16 Architecture générale des mémoires EEPROM.

La figure I.16 est une représentation des principaux circuits composant la mémoire EEPROM. Ces éléments seront détaillés dans cette partie.

a. Génération de la haute tension de programmation

Comme nous l'avons vu, la fonction de mémorisation des mémoires EEPROM nécessite l'application d'un fort champ électrique à travers un oxyde tunnel de manière à obtenir un transfert de charges au niveau de la grille flottante. Et cela, pour des opérations d'écriture comme pour des opérations d'effacement. Ce fort champ est obtenu par l'application de tensions élevées aux bornes de la cellule mémoire.
La production de la haute tension à partir d'une tension d'alimentation est l'un des points les plus critiques : cette haute tension doit, en effet, être stable pour des tensions d'alimentation V_{cc} comprises entre 2.5V et 5.5V, mais aussi dans la gamme de température dans laquelle sont testés les produits (- 40°C à +125°C).
Un cycle de programmation étant composé d'une phase d'effacement suivi d'une phase d'écriture, un signal de programmation double rampe est généré à partir d'un circuit, appelé pompe de charge, interne au produit EEPROM.
Le fonctionnement de la pompe de charge est basé sur un circuit multiplieur de tension de type SCHENKEL [DIC76]. Une représentation simplifiée d'un multiplieur de tension est donnée figure I.17. Ce multiplieur nécessite deux signaux d'horloges déphasés de 180° et non recouvrants. Lorsque le signal Φ est au potentiel V_{cc}, le signal Φ^- se trouve à V_{ss}. Dans ces conditions, les capacités d'indice pair vont se décharger dans les capacités d'indice impair. Puis, lorsque les signaux Φ et Φ^- s'inversent, ce sont les capacités d'indice pair qui se déchargent à leur tour dans les capacités d'indice impair. De cette manière, la tension au nœud d'indice N est augmentée de la valeur V_{in}-V_d par rapport au nœud d'indice N-1 (V_d étant la tension de seuil de la diode).

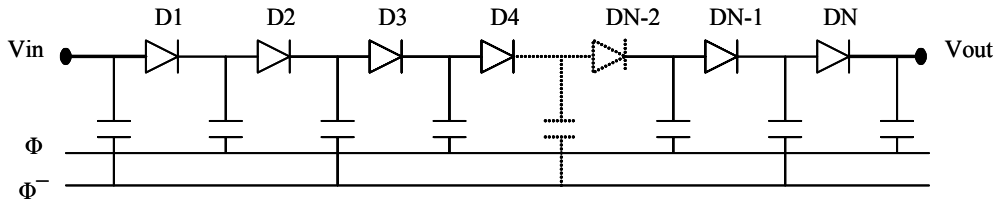


Figure I. 17 Multiplieur de tension.

Une multiplication efficace est réalisée avec des valeurs de capacités relativement hautes. Le multiplieur fonctionne par pompage de paquets de charges le long de la chaîne de diodes, suivant la charge et la décharge des capacités de couplage et cela, à chaque demi-cycle d'horloge. La tension à chaque nœud n'étant pas remise à zéro après chaque cycle de pompage, le potentiel moyen de chaque nœud augmente progressivement du début jusqu'à la fin de la chaîne de diodes.

b. Les opérations de lecture

L'application, sur la grille de contrôle de la cellule mémoire adressée, d'une tension de lecture comprise entre la tension de seuil écrite et la tension de seuil effacée d'une cellule EEPROM permet de différencier les deux états électriques de la cellule. Dans l'état électrique effacé, on sera en présence d'un transistor bloqué. Dans l'état électrique écrit, le transistor mémoire sera dans un état passant.

Dans la pratique, la détection se fait par la lecture du courant traversant le point mémoire. On effectue en général une comparaison entre le courant fourni par une cellule vierge et celui fourni par la cellule à évaluer, toutes les deux polarisées avec une même tension de grille notée V_{ref}. Une conversion courant tension permet ensuite de traduire le résultat de la comparaison en niveau logique. Ce rôle est rempli par un circuit de lecture appelé amplificateur de lecture. Huit amplificateurs de lecture élémentaires montés en parallèle sont nécessaires pour le décodage d'un octet.

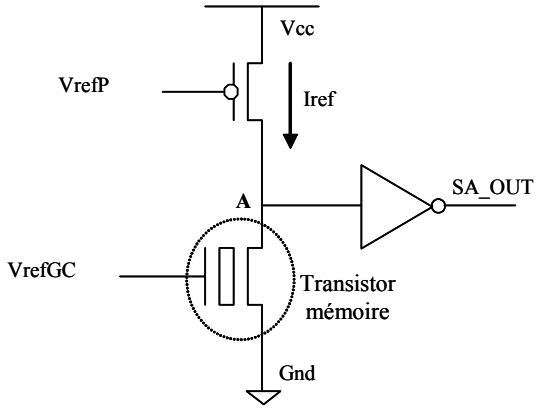


Figure I. 18 Schéma de principe d'un amplificateur de lecture.

Sur la figure I.18, le transistor PMOS est considéré comme un générateur de courant commandé par une tension V_{refP} stable, produite par un organe spécifique (référence de tension). Le courant généré I_{ref} est compris entre le courant parcourant une cellule effacée et celui parcourant une cellule écrite (pour une tension de grille V_{refGC} donnée).
Dans ces conditions, si la cellule à lire se trouve dans un état vierge, le courant parcourant le transistor mémoire est égal à I_{ref}, le potentiel au point 'A' égal à $V_{cc}/2$ et la sortie « SA_OUT » se trouve dans un état indéterminé.
Une cellule effacée ne laisse pas passer le courant de lecture I_{ref} (transistor mémoire bloqué), le transistor PMOS tire le potentiel au point 'A' à V_{cc} et la sortie « SA_OUT » se trouve à l'état bas. En revanche, un transistor mémoire dans l'état écrit laisse passer le courant de lecture I_{ref}, le transistor PMOS tire le potentiel du point 'A' à la masse et la sortie « SA_OUT » se trouve à l'état haut.

c. Interface matrice-circuits périphériques

Tous les signaux utilisés au sein de la matrice mémoire durant les opérations d'adressage, de lecture et de programmation sont acheminés par l'intermédiaire des décodeurs lignes et colonnes. L'acheminement des signaux de programmation nécessite des bascules haute tension encore appelées « latches ». Le rôle des bascules est d'acheminer les signaux de grille et de « bit line » durant les phases d'effacement et d'écriture.
Le principe de fonctionnement d'une bascule est simple (figure I.19) : si la colonne i (signal Col i) est sélectionnée, la sortie du premier inverseur 'A' se trouve au niveau bas, ce qui permet au second inverseur de transmettre les signaux de programmation sur la « bit line ».
Les bascules de « bit line » sont munies en sortie d'un commutateur. Lors d'une opération de lecture, ce commutateur déconnecte la bascule de la « bit line » pour que celle-ci soit reliée à l'amplificateur de lecture.
On retrouve une structure similaire dans le bloc de décodage des lignes de manière à pouvoir appliquer sur la grille du transistor mémoire soit une tension constante lors de la phase de lecture, soit une rampe en tension lors de la phase d'effacement.

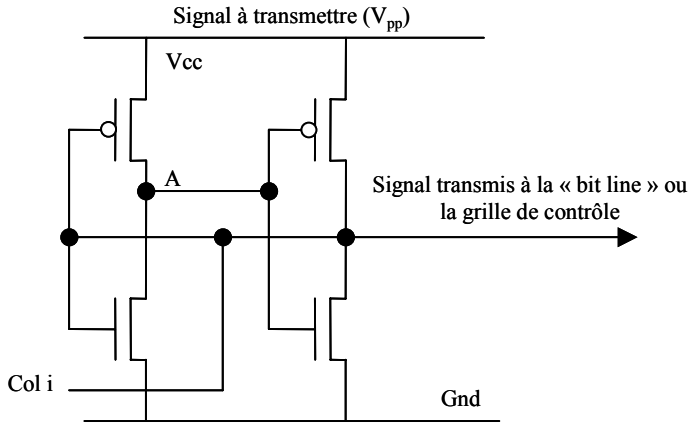


Figure I. 19 Schéma de principe d'une bascule haute tension élémentaire.

d. Gestion de l'information binaire dans le composant

Une logique de contrôle représentée par la machine d'état ou PLA (« Programmable Logic Array ») est chargée de gérer le fonctionnement interne de la mémoire. Le « PLA » se compose d'un plan OU et d'un plan ET. Il est associé à des bascules de type D, ce qui permet de réaliser une machine d'état, qui n'est autre qu'un automate. C'est le cœur du système, il permet de générer les différents signaux internes au circuit durant les phases de lecture et de programmation. En fonction des signaux qu'il reçoit, et selon un organigramme pré-établi, c'est lui qui génère les signaux de commande du circuit.
La logique d'entrée gère le flux d'informations porté sur chaque broche externe du composant. Pour les mémoires disposant d'un protocole de communication série (I^2C, SPI…), les données sont prises en compte successivement sur front d'horloge. Chaque octet est ensuite parallélisé à partir d'un registre à décalage et dirigé vers les décodeurs. Cette opération est réalisée par les registres d'adresses.
Des registres d'état interne au composant et en général accessibles à l'utilisateur, permettent de configurer le mode de fonctionnement de la mémoire et de connaître son état de configuration à tout moment (protection en écriture, processus de programmation en cours…).

II. Le test industriel des circuits VLSI

A. Le test de production

Il est possible d'opposer les tests de type étude qui servent à valider la conception d'un circuit aux tests de type production qui servent à s'assurer que les circuits fabriqués sont garantis dans des limites fixées dans les spécifications du composant considéré. Le test d'étude est un test fonctionnel appelé parfois test architectural par opposition au test structurel qui peut être utilisé en production et qui est basé sur une probabilité de fautes physiques au niveau des portes du composant. Nous nous intéresserons au test de production car il entre directement en compte dans les coûts récurrents des circuits, alors que le test d'étude entre plutôt en compte dans les coûts de recherche et de développement des circuits.

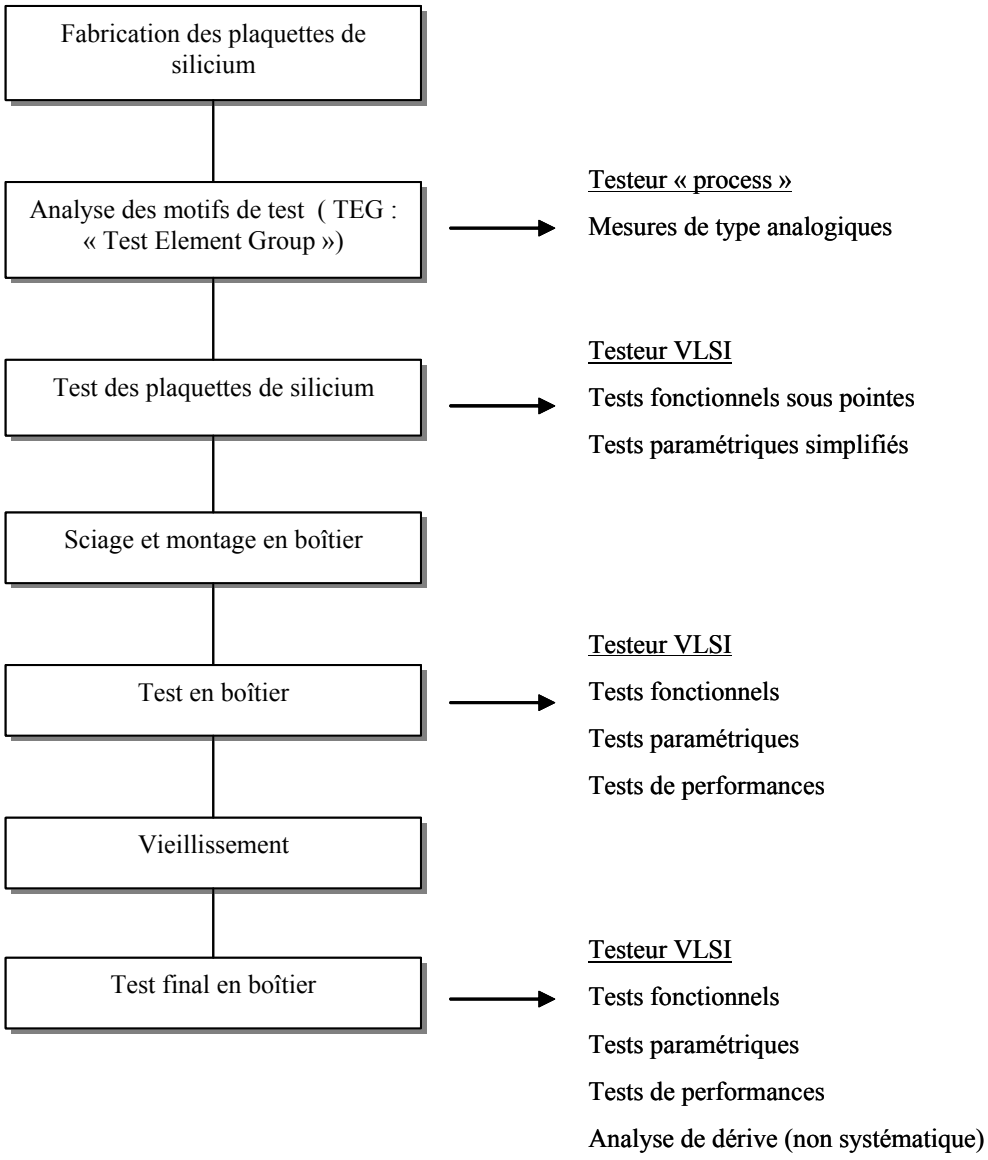


Figure I. 20 Les différents tests lors de la fabrication d'un circuit.

Le test de production suppose que le circuit à tester ou DUT (Device Under Test) ne présente plus d'erreurs de conception et que son architecture soit totalement validée. Ce test doit permettre de vérifier la fonctionnalité de tous les transistors ainsi que l'intégrité de toutes les interconnexions. Pour cela on utilise un outil de génération automatique de vecteurs de test (ATPG : Automatic Test Pattern Generator) qui génère les vecteurs permettant de valider la structure des composants et des interconnexions du circuit. Le but visé est de générer un test assurant une couverture de fautes maximale pour un nombre minimum de vecteurs de tests.
L'organigramme de la figure I.20 présente les différentes étapes de test réalisées tout au long du cycle de fabrication d'une puce électronique.
Tout au long du procédé de fabrication, on distingue principalement trois grandes étapes de test : le test électrique, le test sous pointe et le test en boîtier des puces encapsulées.

1 Le test électrique ou test de motifs (Parametric Test)

Les motifs de test (TEG pour : Test Element Group) sont des éléments qui sont introduits dans la plaquette de silicium par le fondeur. Ils sont placés sur les lignes de découpe

de la plaquette de silicium. Les tests effectués sur ces motifs sont principalement de type analogique. Les motifs sont constitués d'éléments représentatifs du procédé de fabrication (transistors, résistances, capacités, ...). Les mesures effectuées sur ces éléments vont permettre de mesurer des paramètres technologiques tout au long du processus de fabrication.
En général, les motifs de test contiennent des structures permettant de s'assurer de la bonne exécution de chaque étape du processus de fabrication.

2 Le test sous pointe (Probe Test)

Cette étape de test permet de séparer les puces en différentes catégories définies en fonction de leurs performances. Cette opération de tri est appelée « binning ».
Les tests, paramétriques aussi bien que logiques, sont réalisés par l'intermédiaire de testeurs extrêmement coûteux. En conséquence, les testeurs doivent être utilisés au maximum de leurs possibilités. Le test sous pointe est réalisé par un testeur pilotant une machine de test spécifique (appelé « wafer probe » ou encore « prober »). Le testeur est alors associé à un manipulateur qui va prendre en charge les tranches de silicium et le positionnement des pointes devant les contacts. Le test sous pointe est très souvent lié aux unités de fabrication de tranches pour des raisons de propreté d'environnement et de retour d'information rapide en cas de problème.
Cette étape de test est très délicate et nécessite la réalisation de cartes à pointes qui se posent sur les plots de contact du circuit à tester. Ces pointes sont le plus souvent en tungstène. Certains testeurs de taille importante utilisent des cartes à plus de 700 pointes, avec un pas d'espacement de 100 micromètres. Le but étant de tester un maximum de puces en parallèle.
La montée en puissance de la complexité des circuits augmente la difficulté du test, on pratique le plus souvent des tests simplifiés. Par exemple, pour un circuit de type CMOS, on exécutera une série de tests qui comprend l'ensemble des vecteurs de test, plus la mesure des courants de consommation et de fuites sur les entrées. Des tests paramétriques plus poussés se feront au niveau de la puce en boîtier. Le coût très élevé des cartes (au delà de 400 pointes) relance l'intérêt du test en série et plus particulièrement du « Boundary Scan ». Cette technique permet de réaliser un test suffisamment exhaustif en utilisant un nombre limité de pointes.
Toutes les informations relatives au test sont sauvegardées sous un format spécifique à chaque testeur. Les résultats des tests (données logiques et analogiques, « binning »,…) fournissent de précieuses informations aux ingénieurs en charge du test et des produits, qui peuvent les exploiter afin d'améliorer les rendements.

3 Le test en boîtier (Final Test)

Le test en boîtier est à la base du test final qui garantit le bon fonctionnement de la puce encapsulée. A ce titre, il est nécessaire avant toute livraison.
Au cours de ce type de test, nous trouvons les deux catégories de test suivantes :

- les tests paramétriques qui vont garantir les caractéristiques électriques des organes d'entrées-sorties. Ces tests vont vérifier l'aptitude du circuit à dialoguer avec d'autres circuits dans un ensemble logique. Les relevés effectués lors de ces tests permettront de vérifier si le composant respecte les spécifications (data-sheets) du constructeur.
- les tests fonctionnels qui vérifient la fonctionnalité du circuit. Pour cela on applique des vecteurs de test aux entrées du circuit. Le principe de ce type de test consiste à calculer les réponses correspondant à chaque vecteur par simulation à l'aide de divers logiciels, puis de comparer ces réponses avec les mesures effectuées en sortie du composant.

Cependant, un ou plusieurs tests en boîtier pourront être nécessaires selon le type de garantie qualité demandé par le client. Les différents tests associés à ces niveaux de qualité peuvent être des tests de « Burn In » où les boîtiers sont soumis à des cycles de température, des tests de « stress » en tension…
A ce stade du test, les composants à tester sont chargés sur l'unité de test en utilisant un automate de manipulation appelé « handler » qui permet le chargement et le tri des boîtiers.

4 Le test après vieillissement (Burn-In)

Le vieillissement est couramment pratiqué en microélectronique afin d'évaluer la fiabilité d'un composant à long terme. Il consiste à soumettre le circuit ou la puce à une forte température durant un temps donné, avant de le soumettre à de nouveaux tests paramétriques et logiques. Cela est dû à la célèbre courbe en forme de « baignoire » illustrée figure I.21, représentant le taux de défaillance en fonction du temps de fonctionnement du système. Le type de test exécuté ici dépend de la catégorie du composant.
Le test doit garantir un certain niveau de fiabilité après vieillissement. Pour des produits dits à haute fiabilité (militaires ou spatiaux), on peut être amené à faire des corrélations de mesure entre les tests avant et après vieillissement (analyse de dérive). Le but recherché est alors d'assurer que le circuit a atteint une stabilité suffisante. Pour apprécier ces dérives, il faut enregistrer des résultats de mesure et non plus seulement vérifier une fonctionnalité ou une caractéristique par rapport à une limite fixée. Les résultats sont enregistrés dans des fichiers qui serviront à des mesures de dérives ainsi qu'aux corrélations voulues.

Figure I. 21 Distribution des fautes durant la durée de vie d'un produit.

B. Le test et la conception de circuits

1 Défaillances et modèles de fautes

a. Mécanismes de défaillance

Le but du test est de déterminer les circuits défaillants du fait de la présence d'un ou plusieurs défauts (ou défaillances) physiques. Il paraîtrait donc naturel de s'intéresser tout d'abord aux différents types de défaillances pouvant affecter un circuit. Malheureusement, il n'existe pas d'ensemble cohérent de types de défaillances permettant de couvrir tous les défauts pouvant apparaître dans les différentes technologies utilisées pour la réalisation des circuits. Chacune de ces technologies présentant des modes de défaillance bien particuliers.
Les mécanismes de défaillances sont nombreux : ils peuvent être dus à une instabilité du processus de fabrication ou survenir à la suite de dysfonctionnements d'équipements ou après une contamination introduite par un facteur humain [HAY00].

b. Modèles de fautes

La modélisation des défauts est la première étape de la méthodologie de test. Il faut toutefois noter que cette étape n'est pas indispensable car on peut utiliser des modèles de fautes déjà existants. Elle consiste à représenter chaque type de défauts physiques par un modèle de faute défini au niveau électrique ou au niveau logique. L'intérêt de cette représentation est de disposer de modèles facilement manipulables par des outils informatiques.
La définition d'un modèle de faute est basée sur le fait qu'une faute doit provoquer autant que possible le même fonctionnement erroné que l'ensemble des défauts qu'elle modélise. Pour une architecture de circuit donnée, l'application d'un modèle de faute permet de générer une liste de fautes. Ainsi, plus cette liste de fautes se rapproche des défauts réels susceptibles de se manifester, meilleur sera le test.
Un des premiers modèles de fautes à avoir été utilisé est le modèle de collage (« Stuck-at ») appliqué sur un modèle de description du circuit à porte logique. Ce modèle consiste à considérer les collages permanents à 0 et à 1 des différentes lignes de la description logique du circuit.
Malgré de nombreuses améliorations et extensions, des recherches [GAL78] [WAD78] [REN86] ont montré l'insuffisance du modèle de collage pour représenter la majorité des défaillances susceptibles d'être observées. Cette constatation est particulièrement vraie dans les technologies MOS.
Comme nous l'avons mentionné précédemment, une large majorité des défaillances physiques de fabrication est constituée par des circuits ouverts au niveau des connexions et par des courts-circuits entre connexions. De telles défaillances entraînent des erreurs de fonctionnement qui ne peuvent pas être modélisées par un simple modèle de collage qui, de plus, est appliqué à une description à portes logiques.
Certains outils informatiques permettent de déterminer les défauts potentiels pouvant affecter un circuit. Ces défauts sont le plus souvent ramenés à la présence de particules aléatoires (spot defect) qui apparaissent lors des différentes étapes de fabrication du produit. La génération de ces défauts se base uniquement sur la représentation physique (layout) du circuit, ce qui permet d'obtenir une bibliothèque de fautes réaliste.
Le principe de fonctionnement de ces outils est le suivant : la taille des différentes lignes conductrices du circuit est augmentée puis minimisée. Pour chaque variation de la taille des lignes et pour chaque niveau de masque, on évalue de quelle manière le « layout » peut être affecté par des défauts aléatoires [SMF85].

2 Analyse de testabilité

a. Motivation

Concevoir un circuit testable suppose la mise en place d'une véritable stratégie de test dès la conception de l'architecture du circuit. La méthode consiste à établir des règles de testabilité qui sont des règles de conception à part entière. Ces règles dépendent de nombreux facteurs tels que les outils CAO (Conception Assistée par Ordinateur) pour la génération des tests et des moyens de test. La complexité des circuits est devenue telle qu'il n'est généralement pas possible de tester le circuit avec une méthode unique.
L'analyse de testabilité est donc une étape préalable à la génération ou à la simulation de séquences de test. Elle permet d'évaluer les « efforts » nécessaires au test du circuit. La génération des vecteurs de test est une étape extrêmement compliquée et consommatrice en temps CPU. Ceci ne faisant que s'aggraver du fait de l'augmentation constante des fonctionnalités des circuits. Il en résulte une augmentation non négligeable du coût du test et

donc de sa part dans le prix des circuits. L'analyse de testabilité a pour but de définir les performances du test, c'est-à-dire son taux de couverture, le temps de test et la difficulté à tester un circuit donné. Deux paramètres importants sont compris dans la testabilité : la contrôlabilité et l'observabilité.

b. Contrôlabilité

C'est la capacité à forcer un point du circuit à un niveau logique haut ou bas. Elle est maximale pour les entrées mais peut être très réduite pour certains points enfouis du circuit. Il est souvent nécessaire d'appliquer plusieurs vecteurs de test pour avoir accès à un point particulier du composant à tester.

c. Observabilité

Elle représente la facilité à vérifier sur les sorties primaires du circuit la présence d'une valeur logique associée à un noeud. Elle est maximale pour les sorties et est réduite pour les points enfouis du circuit.
La testabilité combine les notions d'observabilité et de contrôlabilité pour quantifier la facilité avec laquelle chacun des nœuds du circuit et donc le circuit dans son ensemble peut être testé. L'analyse de testabilité est principalement réalisée à partir de techniques structurelles basées sur l'analyse d'une description topologique du circuit, mettant en œuvre des portes logiques classiques ou des éléments plus complexes [BRE79] [KOV81].

3 Les différents types de test

Un test de production comporte au minimum trois types de test qui sont : le test DC, le test fonctionnel et le test AC. Il est parfois nécessaire de pré-conditionner le circuit, en utilisant des vecteurs de test spécifiques, avant de réaliser un test particulier.

a. Tests paramétriques DC

Ce sont des tests analogiques qui permettent de déterminer les valeurs des niveaux d'entrée et de sortie du circuit, ainsi que les valeurs de consommation. Ces tests passent par la mesure de plusieurs paramètres :

- les courants et tensions d'entrée (V_{IL}, V_{IH}, I_{IL}, I_{IH}),
- les courants et tensions de sortie (V_{OL}, V_{OH}, I_{OL}, I_{OH}, I_{OZH}, I_{OZL}),
- les courants et tensions d'alimentation (V_{CC}, V_{DD}, V_{DDMIN}, V_{DDMAX}, I_{CC}, I_{DD}).

Ces tests permettent de rejeter ou d'accepter les circuits en fonction des limites définies de ses paramètres DC. Ces limites sont données par les spécifications du constructeur.
Etant donné que les coûts du test sont proportionnels au temps passé, il semble préférable d'éliminer au plus tôt de la chaîne de fabrication les composants défectueux. C'est pourquoi les premiers tests effectués sont les tests éliminatoires les plus rapides.
Le tout premier est le test « opens & shorts ». Ce test permet de détecter rapidement tous les problèmes de broches cassées, de connexions détruites par décharge d'électricité statique ou simplement à cause d'un défaut de fabrication. Les tests DC les plus utilisés sont brièvement décrits ci-dessous :
Les tests V_{OH}/I_{OH} (tension et courant de sortie au niveau haut) : leur but est de vérifier le courant de sortie (I_{OH}) à une tension de sortie donnée (V_{OH}). Ce test mesure la résistance

d'une broche de sortie dans l'état logique haut. Cela permet de vérifier par une mesure de courant, si le système testé délivre suffisamment de courant à une tension donnée.

Les tests V_{OL}/I_{OL} : c'est le même type de test que précédemment mais les mesures de courant sont effectuées en forçant un niveau bas sur la broche.

Le test du courant d'alimentation I_{DD} : ce courant représente le courant circulant de drain à drain dans un circuit CMOS. De la même façon, pour un circuit TTL, on parle de courant I_{CC} qui est le courant circulant de collecteur à collecteur. Ce test est éliminatoire et s'effectue généralement après celui du test des broches afin de rejeter le plus rapidement les composants défectueux. On le retrouve dans le programme de test sous pointe et dans le test de production avant les tests fonctionnels. La mesure du courant I_{DD} s'effectue au niveau de la broche d'alimentation de la puce. Avant la mesure, il est nécessaire de positionner le composant dans un état stable à l'aide d'une série de vecteurs de test.

Les tests I_{DD} statique et I_{DD} dynamique : lors de la mesure en statique, le composant n'est pas actif. Cette mesure est très importante pour connaître la consommation de courant du système et donc prévoir les bonnes capacités de batteries. De plus, une consommation excessive permet de déceler des défauts de fabrication. Comme précédemment, le système doit être positionné dans un état stable par l'application de vecteurs de test. En ce qui concerne la mesure dynamique, le système est soumis à une séquence de vecteurs de tests qui simulent un fonctionnement.

Mesure des courants d'entrée haut et bas I_{IH} et I_{IL} : ces mesures permettent de vérifier la haute impédance des bascules de sortie à haut et bas niveau de tension. Plusieurs méthodes de test sont possibles. La méthode statique série consiste à forcer toutes les broches d'entrées à un niveau de tension bas ou haut (mesure de I_{IL} ou I_{IH} respectivement) sauf celle sur laquelle on applique un niveau opposé et où l'on mesure le courant. On répète l'opération pour chaque broche d'entrée. L'avantage de ce type de test est que l'on mesure chaque broche individuellement et que l'on identifie très facilement les broches défectueuses. L'inconvénient est le temps nécessaire pour tester l'ensemble des broches du circuit.

b. Tests paramétriques AC

Le test AC vérifie que le produit peut réaliser des opérations logiques tout en étant soumis à des contraintes de temps spécifiques.
Ce sont des tests dynamiques qui permettent de déterminer les valeurs de fréquences, de délais, de largeur d'impulsion, de temps de maintien (hold time) et de prépositionnement (set-up time). Ces tests s'appliquent aux signaux d'entrée ou aux signaux de sortie des circuits. Durant ces tests, les conditions d'entrée (V_{IL}, V_{IH}), de sortie (V_{OL}, V_{OH}, I_{OL}, I_{OH}) ainsi que les tensions d'alimentation (V_{DDMIN}, V_{DDNOM}, V_{DDMAX}) sont définies. Ces tests permettent de rejeter ou d'accepter les circuits en fonction des limites définies de ses paramètres AC. Avant le test, il est nécessaire de s'assurer que les performances du testeur utilisé permettent de réaliser les tests définis.

c. Tests fonctionnels

Les tests fonctionnels servent à vérifier le bon fonctionnement interne du circuit, soit en se basant sur des tests structurels [BRE80], sur des tests architecturaux ou bien sur des tests spécifiques, rendus nécessaires par le type d'application (mémoires, circuits analogiques, …).

L'ensemble des vecteurs de test générés par la machine de test est supposé couvrir au mieux l'ensemble des configurations possibles de fonctionnement de tous les blocs implantés sur le circuit. Un optimum en terme de nombre de vecteurs et de couverture de test est à trouver au moment de la génération du programme de test de production. Les difficultés conséquentes à la recherche du test fonctionnel optimal ont débouché sur de nombreuses études, et certaines solutions améliorant l'efficacité du test ont déjà été mises en place.
Ce sont les tests par chemin d'accès (« scan path »), l'auto test ou encore le test embarqué qui peut être appliqué à certaines familles de circuit comme les mémoires par exemple.
Le nombre de vecteurs de test à envoyer vers le circuit à tester va donc être très dépendant non seulement du type de circuit considéré (logique, processeur, mémoire, …), mais également de la conception même du circuit. On voit ici qu'il sera difficile de définir ou plutôt d'estimer le nombre de vecteurs de test nécessaires à un circuit intégré en partant uniquement de sa description fonctionnelle.
Le test fonctionnel doit être exécuté dans toute la gamme de spécification. C'est pourquoi, il est réalisé à V_{DDMIN} et à V_{DDMAX}.
Les broches de sortie à drain ouvert « open drain » peuvent seulement délivrer des tensions de niveau bas. Il est donc nécessaire d'utiliser des moyens externes afin d'atteindre le niveau haut. Par exemple une résistance ou bien un courant dynamique de charge.
A l'opposé, les broches de sorties source ouverte « open source » peuvent uniquement délivrer des tensions de niveau haut. De la même manière, on peut sortir des niveaux bas par une aide externe. Lorsque le circuit fonctionne à haute fréquence, et qu'une broche drain ouvert passe d'un niveau bas à un niveau haut, le temps de commutation dépend du courant de charge externe et de la capacité associée. Lors de l'exécution du test fonctionnel, si une résistance est choisie trop grande ou bien un courant de charge trop faible alors le temps de changement de niveau peut ne pas être suffisamment rapide.

d. Le test et les gammes de produits

Différents types de tests peuvent être réalisés selon les conditions de température ou les besoins en terme de qualité de composants.
Les contraintes de qualité définissent trois grandes gammes de composants électroniques. Dans l'ordre croissant d'exigence on trouve : la gamme commerciale, la gamme industrielle et la gamme militaire et automobile (Tableau I.1).

GAMME COMMERCIALE	La tolérance est de **100 ppm**. Les tests effectués sont calculés pour juste atteindre cette limite.
Test Probe	test à 25 °C, durée de 3 à 10 sec (sans optimisation)
Test Final	test à chaud (150°C)
GAMME INDUSTRIELLE	Tolérance un peu plus stricte
Test Probe	idem
Test Final	test à chaud + test froid (25°C) ou encore appelé stress
GAMME AUTOMOBILE ET MILITAIRE	La tolérance n'est que de **10 ppm**, cela implique un accroissement important du nombre de vecteurs de test, environ dix fois plus.
Test Probe	Test complet et sauvegarde des résultats de mesures
Test Final	les tests sont effectués à chaud (5 sec), à froid (5 sec) et à 25 °C (5 sec). Ils sont précédés d'un Burn-In (vieillissement accéléré) de 12 heures à 150°C.

** 1 ppm : Une partie par million, soit ici 10 à 100 composants défectueux par million*

Tableau I. 1 Le test et les gammes de produits.

e. L'ATPG (Automatic Test Pattern Generation)

La manière traditionnelle pour tester un circuit logique numérique est de générer des vecteurs de test. L'ensemble de ces vecteurs de test est généré à l'aide d'outils logiciels de type ATPG (Génération Automatique de Vecteurs de Test) qui peuvent en un temps restreint produire l'ensemble des vecteurs requis pour assurer, à un coût minimum, le bon fonctionnement des circuits intégrés.
Il s'agit d'appliquer une série de combinaisons logiques binaires aux entrées du circuit. Pour chaque stimulus, on vérifie le résultat obtenu aux bornes du circuit par comparaison aux calculs. Le problème du test des circuits à très haut niveau d'intégration est qu'il est impossible d'appliquer toutes les combinaisons binaires dénombrables aux bornes du circuit. La sélection des combinaisons utiles est un des problèmes fondamentaux du test. Le véritable problème de ce test est l'évaluation de son exhaustivité (rapport du nombre de fautes détectées sur le nombre de fautes totales) [WAI90]. On admet généralement que pour obtenir un test exhaustif il faut trois vecteurs de test par transistor implanté sur le circuit.

Cette brève description des différents types de test montre que le test d'un circuit intégré va dépendre de nombreux paramètres. Ces paramètres sont liés soit au nombre et au type d'entrées-sorties, soit au nombre de transistors ou de portes implantés sur la puce, soit à la gamme du produit testé, ou encore, au type d'application.

4 Le test intégré ou BIST (Built In Self Test)

Le test d'un circuit nécessite d'appliquer des séquences de vecteurs à ses différentes parties et d'observer les réponses afin de pouvoir les comparer aux réponses attendues. Pour cela, il est supposé implicitement que ces vecteurs sont appliqués sur le circuit par l'intermédiaire d'un élément extérieur (qui, dans la grande majorité des cas est un testeur). L'utilisation d'un testeur extérieur se trouve confrontée à un certain nombre de problèmes dont les plus cruciaux sont les suivants :

- le prix élevé et toujours croissant des testeurs,
- la difficulté et le temps nécessaire pour générer les séquences de test,
- le temps très élevé nécessaire pour appliquer de longues séquences de test,
- la perte relative de puissance des testeurs (en terme de vitesse, nombre de broches envisageables, ...) vis à vis des circuits à tester.

Afin de diminuer ces problèmes, des techniques connues sous le nom de test intégré ont été proposées afin d'inclure directement dans le circuit tout ou partie des fonctions réalisées par le testeur. Comme un testeur extérieur, les dispositifs de test intégré devront donc être capables de générer des vecteurs de test et de comparer les résultats obtenus à ceux attendus suivant le schéma de principe représenté figure I.22.

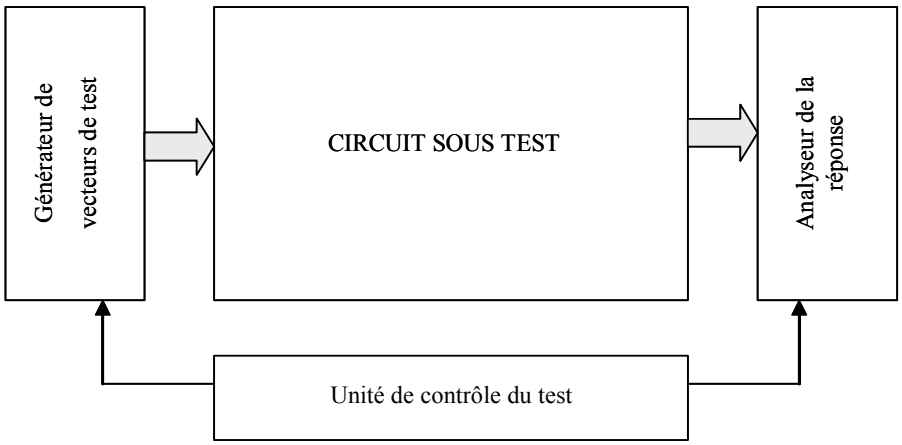


Figure I. 22 Technique de test intégré (BIST).

Le BIST est un acronyme IEEE (inventé en 1982) qui représente la fusion des termes « Built in Test » et « Self Test », en d'autres termes la fusion du test intégré et de l'auto-test.
Le test intégré présente de ce fait les avantages suivants :

- suppression de la nécessité d'un testeur coûteux,
- possibilité de test à haute fréquence et à la vitesse nominale de fonctionnement du circuit,
- taux de couverture élevé,
- temps de test court en utilisant à la fois le parallélisme et la hiérarchie du circuit, avec une vitesse de test égale à la vitesse nominale du circuit,
- possibilité de tester le circuit en phase de maintenance mais aussi en opération.

En contrepartie, l'utilisation du test intégré conduira obligatoirement à un surcoût en matériel, s'accompagnant d'une perte de performances.

C. Les machines de test

L'introduction des notions essentielles du test utilise un jargon de test assez déroutant au premier abord, composé d'une collection d'acronymes et de termes anglais. Afin de faciliter la compréhension du lecteur, un glossaire est disponible en annexe de ce rapport.

1 Vue générale

De manière générale le test d'un circuit s'effectue durant toutes les étapes de sa fabrication par l'utilisation de tests paramétriques aussi bien que fonctionnels.
Le système de test est constitué d'une partie matérielle électronique et mécanique utilisée pour simuler les conditions de fonctionnement auxquelles le produit à tester sera soumis au cours de ses futures applications. Les systèmes de test sont souvent désignés par l'acronyme ATE (Automated Test Equipement). Durant les tests sous pointes, les tranches de silicium sont testées en utilisant des cartes à pointes. Le testeur étant associé à un manipulateur nommé « prober » qui va prendre en charge la manipulation des tranches. Afin de minimiser les coûts, le test sous pointe sera le plus complet possible afin d'éviter le montage de pièces mauvaises. Le test en boîtier est à la base le test final qui garantit le bon fonctionnement de la puce encapsulée. Dans ce cas, les composants à tester sont chargés sur l'unité de test en utilisant un automate de manipulation nommé « handler » qui permet le chargement et le tri des puces en boîtiers.

2 Architecture d'un testeur

La figure I.23 présente les blocs fondamentaux qui composent tout système de test digital. Ce diagramme qui constitue une solide référence, ne présente néanmoins pas l'ensemble des éléments matériels contenus sur les testeurs les plus récents.
Je me limiterai donc dans cette partie à la présentation des principaux composants des testeurs.

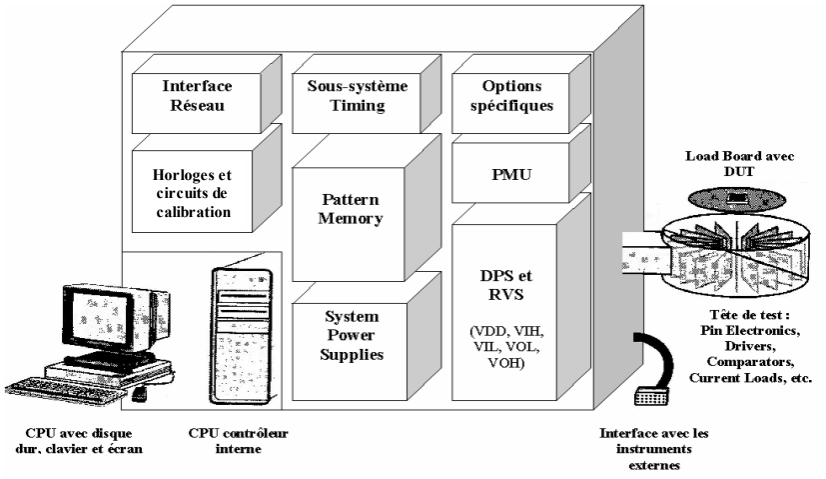


Figure I. 23 Architecture d'un testeur.

a. Le contrôleur interne (CPU)

Le contrôleur du système (ordinateur) contrôle les éléments du testeur et fournit le moyen de transférer des instructions au testeur. La plupart des nouveaux systèmes de test proposent une interface réseau.

b. La partie matérielle

Parmi les composants fondamentaux constituant la partie matérielle, on distingue les éléments suivants :
Un sous-système DC : il contient les alimentations (DPS) et les tensions de référence (RVS) nécessaires pour fournir les niveaux logiques hauts et bas. Ces tensions sont représentées par les symboles V_{IL}, V_{IH}, V_{OL} et V_{OH}. Cette partie contient aussi l'unité de mesure de précision appelée PMU (Precision Measurement Unit).
Cette unité de mesure est utilisée pour effectuer des mesures de tension et de courant précises. Elle peut forcer un courant et mesurer une tension ou bien forcer une tension et mesurer un courant. Une sélection d'échelle de mesure appropriée doit assurer les résultats de test les plus précis possibles.
L'unité de mesure de précision possède deux limites de mesure programmables, une limite supérieure et une limite inférieure. Ces limites peuvent être utilisées individuellement ou simultanément. Si la valeur mesurée n'appartient pas à ces intervalles, le test échoue. Si la valeur mesurée est comprise dans les limites de test définies, le test est réussi.

Une mémoire de vecteurs de test : chaque système de test possède une mémoire à accès rapide (« Pattern Memory ») pour stocker les vecteurs de test. Les vecteurs de test contiennent les états des entrées et des sorties pour les différentes fonctions logiques que le produit doit

remplir. Au cours du test, les groupes de vecteurs (ou « patterns ») sont appliqués sur la puce par le système de test et les signaux de réponse sont capturés. Si la réponse attendue ne correspond pas à la réponse issue du produit testé, un échec fonctionnel survient.

Le sous-système temporel (ou sous-système « Timing ») : il possède une mémoire qui permet de stocker des données temporelles afin de les utiliser durant les tests fonctionnels. Le sous-système reçoit les vecteurs d'entrée provenant de la mémoire de vecteurs de test et les combine aux informations temporelles pour générer des signaux formatés qui sont envoyés à la puce.

La tête du testeur : elle contient les cartes qui renferment les canaux du testeur, ainsi que les interfaces matérielles avec les produits testés. Les canaux (ou « channels ») sont les circuits situés dans la tête de test et qui permettent d'appliquer ou de lire les tensions et les courants sur les broches du produit testé. L'application de ces signaux se fait au travers d'une carte spécifiques (« Pin Electronics »).

La « Pin Electronics » (également appelée « Pin Card », « PE », « PEC » ou « I/O Card ») : elle constitue l'interface entre les ressources du système de test et la puce. Elle fournit les signaux d'entrée au produit testé et en reçoit les signaux de sortie. Le testeur comprend autant de « Pin Electronics » que de canaux.

La carte d'interface produit : elle constitue l'interface physique entre le système de test et le produit testé. Elle contient les composants de l'interface (relais, résistances ou capacités) nécessaires au test de la puce. Elle est également désignée par le sigle DIB (Device Interface Board).

c. La partie logiciel : le programme de test

L'objectif du programme de test est de contrôler le test matériel de manière à garantir que le produit testé réponde à tous ses paramètres de production. Ces paramètres sont définis dans les spécifications du produit. Le programme de test est souvent segmenté en plusieurs parties telles que les tests DC, les tests fonctionnels ou les tests AC dont je rappelle les rôles :

- le test DC vérifie les paramètres en tension et en courant,
- le test fonctionnel s'assure du bon fonctionnement des diverses fonctions logiques du produit,
- le test AC vérifie que le produit peut réaliser des opérations logiques tout en étant soumis à des contraintes de temps spécifiques.

Le programme de test sépare les produits testés en différentes catégories définies en fonction de leurs performances. Cette opération de tri est appelée « binning ».
Le programme de test doit également être capable de contrôler les éléments matériels externes comme les manipulateurs (« handlers » ou « probers »). De plus, le programme de test prend en charge la collecte et la gestion les résultats des tests.

3 Les paramètres importants d'un testeur

Voici les quelques paramètres importants lors du choix d'un testeur. Ce sont principalement eux qui influent sur le prix d'achat. Parmi ces paramètres, on distingue :
Le nombre de canaux du testeur : ce nombre de canaux doit être égal au nombre de broches logiques du testeur. On trouve aujourd'hui des testeurs à 128, 256, 512 et même 1024

broches. Si le testeur est destiné au test parallèle il doit posséder une unité de mesure de précision par broche.

La fréquence de fonctionnement maximale : elle doit être supérieure à la fréquence de fonctionnement du circuit pour le test de mesure de vitesse de fonctionnement maximum. Globalement, plus le testeur est rapide et plus il est coûteux.

La précision sur les définitions temporelles : cette information est souvent plus importante que la fréquence de test. C'est la précision avec laquelle nous allons pouvoir construire le chronogramme temporel de fonctionnement du circuit. C'est aussi la précision des mesures temporelles. Cette précision intervient dès que l'on veut réaliser les mesures destinées aux spécifications du composant (i.e. caractériser tous les paramètres DC et AC présents dans les spécifications). La précision obtenue avec les testeurs haut de gamme est de l'ordre de la picoseconde. Pour obtenir ce résultat, il faut utiliser des systèmes de correction et d'auto calibrage capable de prendre en compte le temps de propagation des signaux entre la station de test et le circuit.

La capacité du testeur en nombre de vecteurs de test : ce paramètre est lié à la taille de la mémoire locale du testeur et à la puissance informatique de l'ordinateur pilote. C'est souvent un facteur limitant pour la rapidité d'exécution des tests car le testeur doit charger les vecteurs de test dans sa mémoire interne en plusieurs fois.

III. Le test des mémoires

Les systèmes digitaux de type microprocesseurs, ASIC, ou SOC (« System On Chip ») embarquent de plus en plus de mémoire. Cette dernière possède, relativement à la partie logique, un taux assez élevé de défauts. Ceci est dû d'une part à la surface importante occupée par la mémoire et d'autre part à sa densité d'intégration élevée.
Les mémoires présentent des propriétés de régularité remarquables. Dès lors, la détection et la localisation de défauts au sein d'un plan mémoire est d'une importance capitale pour la qualité de la mémoire elle-même, mais aussi pour celle du système tout entier.
On conçoit donc qu'il sera intéressant, plutôt que de faire appel aux méthodes générales mises au point pour les circuits séquentiels, d'utiliser les spécificités des mémoires, c'est-à-dire essentiellement leur fonction, et la régularité de leur structure, pour élaborer des méthodes mieux adaptées. Et cela, en particulier au niveau de la modélisation des fautes que l'on cherche à détecter.
Il existe un très grand nombre de procédures de test fonctionnel des mémoires [VAN91], qui se caractérisent par leur taux de couverture et leur durée. Ces procédures ciblent particulièrement les mémoires vives.
Les mémoires non volatiles de type EEPROM se démarquent de ce type de mémoires, de par leur architecture et leur mode de fonctionnement. Il en résulte des procédures de test adaptées.

A. Test des mémoires vives

La durée de test de mémoires vives est généralement représentée par l'expression de son ordre de grandeur en fonction de n [VAN91], où n représente la complexité de l'algorithme de test. Ainsi, les procédures dites traditionnelles qui étaient utilisées avant les années 80 étaient du type $nLog_2n$, $n^{3/2}$ voire n^2. Ces procédures donnaient satisfaction au

regard des retours client compte tenu de leur temps d'application. Avec l'apparition de mémoires de capacités de plus en plus importantes, ces méthodes ne sont plus économiquement viables au vu des temps de test.

1 Modélisation fonctionnelle et modèles de fautes [LAN99].

La figure I.24 donne une représentation fonctionnelle générale d'une mémoire DRAM (RAM dynamique), pour une mémoire SRAM, la logique de rafraîchissement serait absente. Cette modélisation fait apparaître une logique de décodage constituée du registre d'adressage et des décodeurs lignes et colonnes ainsi qu'une logique de lecture et d'écriture qui comprend l'amplificateur de lecture, le registre de données ainsi que les bascules d'écriture. Les signaux de lecture et d'écriture sélectionnent le mode de fonctionnement de la mémoire.

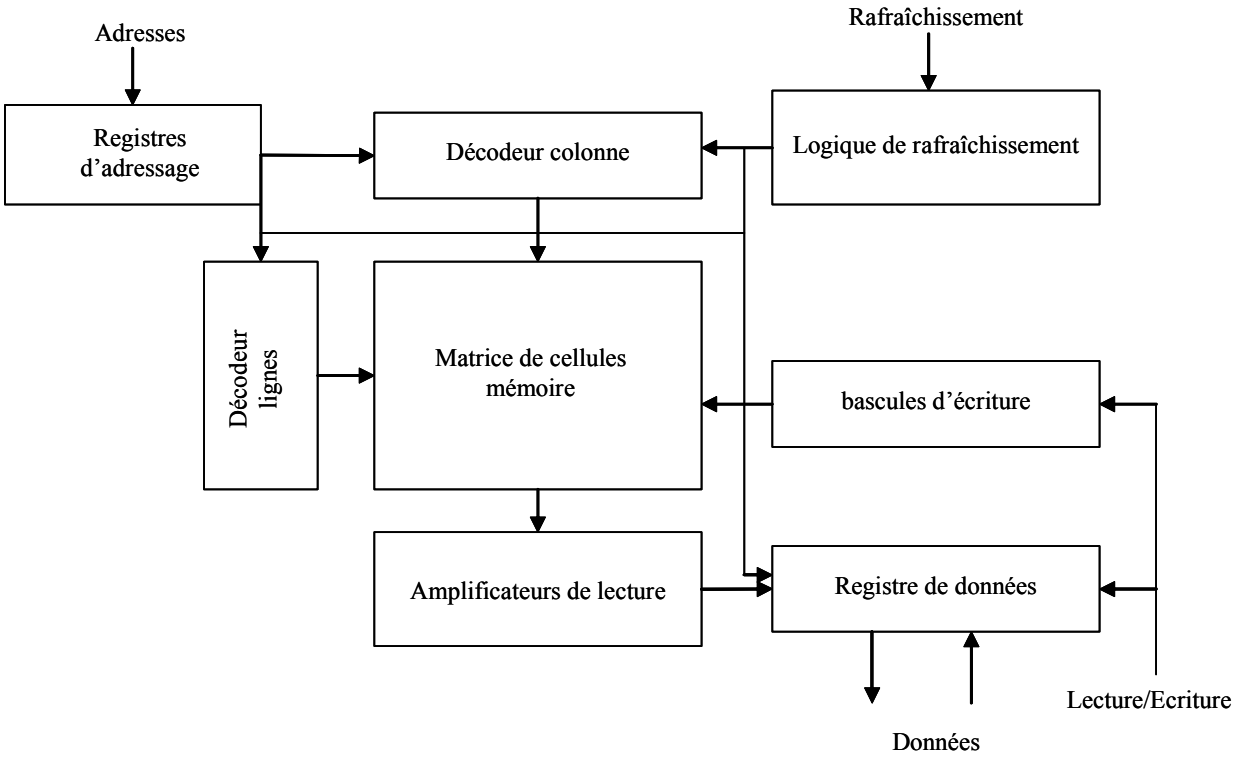


Figure I. 24 Schéma bloc fonctionnel d'une mémoire RAM dynamique.

2 Fautes fonctionnelles [LAN99].

Suivant la manière dont elles se manifestent, les fautes sont classées en fautes simples et en fautes liées :

- une faute simple est une faute qui n'influence pas le comportement des autres fautes,
- les fautes liées influencent le comportement d'autres fautes de telle manière que des masquages peuvent apparaître [PAP85] [VAN90]. Elles ne seront pas traitées ici.

Les fautes simples peuvent être subdivisées en fautes d'adressage, liées au décodeur d'adresse et en fautes mettant en cause la matrice de cellules mémoires. Ces dernières peuvent affecter une cellule du plan mémoire ou plusieurs cellules à la fois.

a. Fautes affectant le décodage des adresses

On distingue quatre types de fautes affectant les décodeurs d'adresses :

- avec une certaine adresse aucune cellule n'est accédée,
- il n'y a pas d'adresse avec laquelle une cellule particulière peut être accédée,
- avec une certaine adresse, plusieurs cellules sont accédées simultanément,
- une cellule particulière peut être accédée avec plusieurs adresses.

b. Fautes affectant les cellules du plan mémoire

Les fautes simples de cellules sont restreintes à une seule cellule. Cette classe de fautes comprend les types suivants :

- les fautes de collage SAF (Stuck-At-Fault). La valeur logique d'une cellule collée est toujours 0 ou toujours 1,
- les fautes de collage ouvert SOF (Stuck-Open-Fault). Ce type de faute dans une cellule mémoire signifie que la cellule ne peut être accédée. Lorsque l'amplificateur de lecture contient une bascule alors une opération de lecture peut fournir la valeur lue précédemment,
- les fautes de transition TF (Transition-Fault). Une cellule n'arrive pas à effectuer la transition de l'état logique 0 vers l'état logique 1 ou/et une transition de l'état logique 1 vers l'état logique 0,
- les fautes de rétention de données DRF (Data Retention Fault). Dans ce cas une cellule n'arrive pas à maintenir sa valeur logique après une certaine période de temps. Dans ce cas, soit un état logique 1 mémorisé devient un 0 après un temps donné, soit un état logique 0 mémorisé devient un 1.

Les fautes entre cellules mémoire mettent en jeux au moins deux cellules du plan mémoire. Les fautes de couplage affectent deux cellules quelconques du plan mémoire : une cellule C_j sensibilisant la faute dans une autre cellule C_i. La cellule C_j est appelée la cellule couplante et la cellule C_i la cellule couplée. On distingue plusieurs types de couplage :

- Les fautes de couplage d'inversion (C_{Fin}). On dit que la transition de l'état logique 0 vers 1 (ou 1 vers 0) de la cellule C_j a une influence d'inversion sur la cellule C_i si cette dernière change d'état.
- Les fautes de couplage idempotentes (C_{Fid}). On dit que la transition de l'état logique 0 vers 1 (ou 1 vers 0) de la cellule C_j a une influence idempotente à 1 sur la cellule C_i si elle la fait passer de 0 à 1, mais pas de 1 à 0 ou en d'autres termes si elle force la cellule C_i à 1.
- Les fautes de couplage d'état (C_{Fst}), dans ce cas une cellule couplée C_i est forcée à une certaine valeur logique 'x' seulement si la cellule couplante C_j est dans un état donné 'y' [DEK90].
- Les fautes de perturbation (C_{Fdst}) sont des fautes par laquelle la cellule couplée est perturbée (change d'état logique) à la suite d'une lecture ou d'une écriture de la cellule couplante [DEJ76].

Le cas général des fautes de couplage de plusieurs cellules est extrêmement complexe à traiter si l'on ne pose aucune restriction sur la position respective de ces cellules. Si l'on se limite par contre aux cellules voisines d'une cellule de base donnée on parle alors de fautes sensibles à la configuration des cellules voisines ou fautes de voisinage.

3 Procédures de test de type n

Il y en a de nombreuses, qui consistent, par exemple, à placer une même donnée dans toute la mémoire (ce qui ne permet pas de tester le système d'adressage), ou à écrire des données complémentaires dans des cellules topologiquement adjacentes (« checkerboard »), pour tester les courts-circuits entre ces cellules (ce qui permet de tester les erreurs sur le bit de moindre poids d'adressage de ligne ou de colonne, mais pas les autres). Ces procédures consistent aussi à écrire des données complémentaires dans des colonnes (« column bars ») ou lignes (« row bars ») adjacentes, ou à écrire sur la diagonale des données complémentaires de celles des autres cellules (« diagonal »), ou à écrire dans chaque cellule le bit de parité de son adresse (« parity pattern »). Les taux de couverture associés à ces procédures de type n sont faibles et seule la détection des fautes de collage est garantie [VAN91].

4 Procédures de test de type n^2

a. Le un (zéro) baladeur (« Walking 1 (0) »)

Ce type de test débute par l'initialisation de toute la mémoire à l'état logique 0 (1), puis on écrit la donnée logique 1 (0) dans la première cellule. On lit alors toutes les autres cellules pour vérifier qu'elles sont bien à l'état 0 (1) et la première cellule pour vérifier qu'elle est bien dans l'état 1 (0). On écrit ensuite un 0 (1) dans la première cellule et un 1 (0) dans la seconde cellule, puis on lit toutes les cellules pour vérifier qu'elles sont bien à l'état 0 (1) exceptée la seconde qui est à l'état 1 (0). La séquence est répétée jusqu'à ce que chaque cellule de la mémoire ait été positionnée à 1 (0). Cette méthode permet de prouver que chaque cellule peut être mise à l'état 0 ou 1 sans modifier l'état des autres cellules et que l'adressage s'effectue correctement.

b. Le test « GALPAT » pour « GALoping PATtern »

Il s'agit d'une extension de la méthode précédente qui permet de tester toutes les transitions d'adresse, ce qui rend possible, quand on ne connaît pas la logique interne de décodage, de tester les transitions de pire cas (entre autres). L'idée est la suivante : après avoir écrit la donnée logique 1 (0) dans une cellule, au moment de vérifier qu'il y a toujours des 0 (1) dans les autres cellules, on fait précéder les ordres de lecture de chacune des autres cellules par la lecture de la cellule qui vient d'être modifiée. Cette deuxième méthode requiert donc approximativement deux fois plus d'opérations que la méthode du zéro baladeur, ce qui ne change pas l'ordre de grandeur, qui reste n^2.
Les procédures de type n^2 ont des taux de couverture excellents, par contre elles présentent des temps d'exécution prohibitifs pour les mémoires de capacité importante.

5 Procédures de test de type $n^{3/2}$

De telles méthodes ont été proposées pour tenter de réaliser un compromis entre les procédures de type n, dont le taux de couverture n'est pas excellent, et les procédures de type n^2 qui, comme nous l'avons vu, requièrent un temps excessif pour des mémoires de capacité importante. Par exemple, un test industriel a été proposé [BAR76], basé sur le principe du zéro baladeur, mais en opérant successivement sur chaque colonne, considérée comme indépendante des autres (test connu sous le nom de « GALCOL »). De la même manière, on trouvera un test n'agissant que sur chaque ligne (test connu sous le nom de « GALROW »).

6 Les tests « MARCH »

Les tests les plus simples et les plus efficaces qui permettent de détecter les fautes de collages, les fautes de transition et les fautes de couplage appartiennent à une catégorie de test appelée MARCH. Un test de type MARCH est un test tel qu'il parcourt toutes les cellules mémoires une ou plusieurs fois par ordre croissant ou décroissant en appliquant successivement sur chaque cellule une séquence d'opérations données appelée élément. Les séquences d'opérations peuvent consister en :

- écrire un 0 dans une cellule (w0),
- écrire un 1 dans une cellule (w1),
- lire une cellule avec une valeur 0 attendue (r0),
- lire une cellule avec une valeur 1 attendue (r1).

Chaque élément est regroupé entre parenthèses, les opérations étant séparées par des virgules. Par exemple l'élément (r0, w1) signifie lire la cellule en espérant trouver la valeur 0 puis écrire un 1. Chaque élément est précédé du symbole ↑ pour un parcours croissant des adresses, du symbole ↓ pour un parcours décroissant des adresses ou du symbole ↕ si l'ordre des adresses est indifférent. Les différents éléments sont séparés par un point virgule et le test complet est représenté entre accolades.
Ainsi la notation abrégée {↑(r0,w1);↓(r1,w0)} représente un algorithme qui va, dans un premier temps lire les cellules dans l'ordre croissant des adresses en attendant la valeur 0 et dans le même temps écrire la valeur 1. Puis, relire la mémoire dans l'ordre décroissant de ses adresses en attendant la valeur 1, en y écrivant la valeur 0.
Il a été montré que toutes les fautes de décodage d'adresse, de collage, de collage ouvert ainsi que toutes les fautes de transition peuvent être détectées par une séquence spécifique de test MARCH [VAN90] [VAN93]. Tout test MARCH peut être étendu afin de détecter les fautes de rétention de données [VAN90] [VAN93]. Il suffit que chaque cellule soit placée dans les deux états possibles et qu'un délai suffisant soit ensuite respecté.
L'ouvrage de Christian LANDRAULT : "Test, testabilité, et test intégré des circuits intégrés logiques" présente une synthèse des tests MARCH les plus connus, ainsi que les types de fautes couverts par ces différents algorithmes de test [LAN99].

7 Test intégré des mémoires vives

L'utilisation massive des circuits mémoire dans les circuits actuels peut poser des problèmes de test sérieux car très souvent les lignes de données, d'adressage et de contrôle des blocs mémoires ne sont pas directement accessibles au travers de broches extérieures. Le test intégré a de ce fait été envisagé très tôt pour ce type de circuit.
Bien qu'il ait été proposé d'utiliser un test pseudo-aléatoire pour le test des mémoires vives [DAV89] [MAZ92], il convient ici de mettre en avant la solution utilisant un test déterministe basé sur l'application des tests March décrits précédemment. En effet, la régularité des tests March les rend tout particulièrement adaptés à une utilisation dans le test intégré.

B. Test des mémoires non volatiles

1 Spécificité des mémoires non volatiles

a. Organisation et fonctionnement des mémoires non volatiles de type EEPROM

L'architecture d'une mémoire non volatile est globalement semblable à celle d'une mémoire vive. Elle comprend un plan de cellules mémoire, chacune des cellules étant identifiée par son adresse. Ce plan est entouré, de manière classique par les éléments suivants (figure I.25) :

- une logique de commande (PLA, « Programable Logic Array ») avec des registres de contrôle,
- un décodeur de ligne pour la sélection d'une ligne active,
- un décodeur de colonne (ou de « bit line ») pour la sélection de la colonne active,
- une circuiterie de lecture dont les niveaux de lecture sont fixés par des références en tension et courant.

On note par ailleurs que l'architecture de la mémoire présente des spécificités directement liées à son mode de programmation qui utilise des niveaux de tension élevés (V_{pp}). Les éléments spécifiques aux mémoires non volatiles sont les suivants :

- un étage de génération des signaux de programmation haute tension nécessaires durant les phases de programmation du plan mémoire,
- les bascules haute tension dont le rôle est de mémoriser les données à écrire dans chaque cellule du plan mémoire et d'acheminer la haute tension durant la phase de programmation,
- une logique de décodage des lignes conçue dans le but de transmettre la haute tension sur des lignes actives durant la phase de programmation,
- un étage de redondance qui va substituer les lignes ou colonnes contenant des cellules mémoires défaillantes par des lignes ou colonnes « saines ». Ces dernières se situent sur les côtés de la mémoire et sont adressées, en fonction des résultats des tests fonctionnels, par le bloc de redondance.

Le test des mémoires non volatiles de type EEPROM nécessite la prise en compte des spécificités de ce type de mémoire. Je rappelle qu'une opération de mémorisation de données passe par une phase d'effacement (stockage d'électrons dans la grille flottante des cellules adressées) et une phase d'écriture (destockage des électrons). Pour cela, l'architecture de la mémoire inclut dans sa puce un générateur de haute tension, telle qu'une pompe de charge. Cette pompe de charge comprend un oscillateur (optimisé pour être stable en fréquence) qui attaque un multiplicateur de tension de type Schenkel que nous avons décrit dans la première partie du chapitre [DIC76]. La double rampe de programmation est générée par le générateur de rampe. La figure I.25 représente un schéma bloc fonctionnel d'une mémoire non volatile de type EEPROM. Les blocs de couleur sombre sont liés à la génération ou l'acheminement de la haute tension.
Des modes de défaillance particuliers, liés d'une part à l'architecture de la mémoire et d'autre part à son mode de fonctionnement, devront être pris en compte lors d'une analyse de défaillance ciblant ce type de produit. En effet, les technologies mémoires non volatiles à grille flottante nécessitent des tensions de programmation assez élevées pour être opérationnelles. Par conséquent, il existe au sein de ces structures des champs électriques relativement forts, notamment aux bornes des oxydes de grille. De plus, dans le cas de mémoires embarquées (mémoires insérées dans un circuit logique conçu dans le procédé de fabrication CMOS classique), il a fallu introduire de nouvelles méthodes de conception et de fabrication. Dans cette technologie de fabrication mixte (logique et mémoire non volatile), il existe trois zones distinctes, fabriquées de manière différente. La première correspond au cœur mémoire. La seconde est une zone appelée « zone haute tension » : elle est composée de transistors ayant une épaisseur d'oxyde assez élevée pour résister à la haute tension de

programmation. Enfin la dernière zone est qualifiée de « zone basse tension » et représente la partie CMOS.

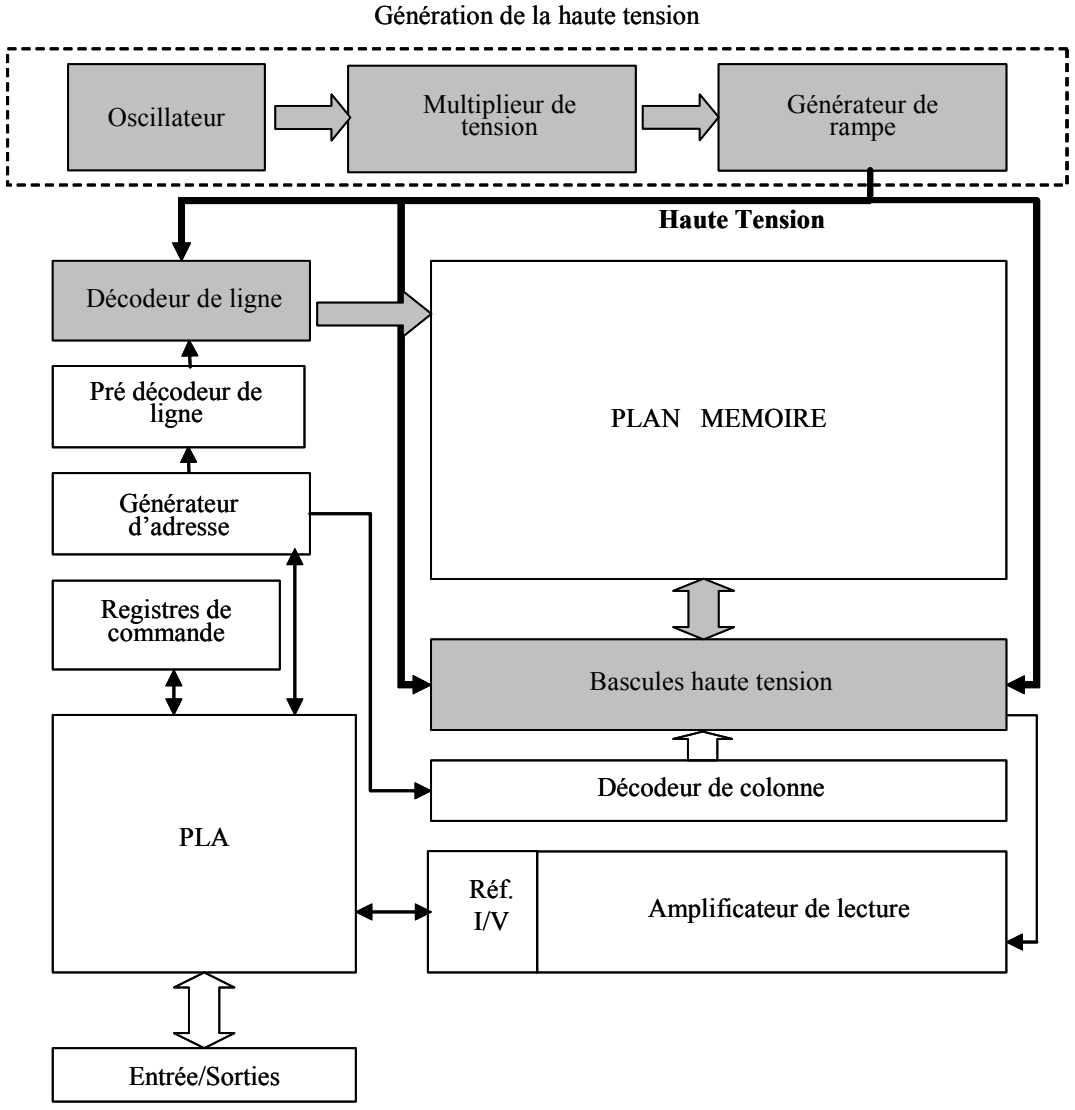


Figure I. 25 Architecture d'une mémoire non volatile de type EEPROM.

b. Fiabilité et modes de test spécifiques aux mémoires EEPROM

La fiabilité des mémoires non volatiles repose principalement sur trois critères : la fonctionnalité, la rétention et l'endurance. Ces trois caractéristiques devront être validées durant les phases de test.
A cela s'ajoutent des critères de performance incontournables comme le temps d'accès en lecture, la durée de programmation, la consommation en mode de lecture, d'écriture ou de veille, la surface occupée…
La mémoire peut être utilisée en mode fonctionnel (repos, lecture, écriture de données…) ou en mode test suivant la configuration des registres de commande. Le mode test permet d'évaluer les performances de la mémoire et donne la possibilité d'effectuer certaines caractérisations. Le mode test comprend plusieurs types de lecture et d'écriture.

- le mode page, dans ce mode plusieurs mots de la mémoire peuvent être programmés en un cycle sur toute une ligne. Ce mode de test est aussi utilisé durant la phase de lecture pour les mémoires à accès série,

- le mode global, qui correspond à la sélection d'une colonne de la mémoire avec le mode page actif. Ce mode de test permet l'écriture de tout le plan mémoire avec une certaine donnée,
- l'accès sectorisé à la mémoire, dans ce cas les opérations de programmation et de lecture s'effectuent sur une partie spécifique de la mémoire (demi-plan, quart de la mémoire, huitième de mémoire….),
- l'extraction des tensions de seuil, qui permet d'évaluer, dans un premier temps, la fonctionnalité de la mémoire (fenêtre de programmation). Ce mode de test est aussi utilisé pour contrôler le décalage des tensions de seuil de toutes les cellules du plan mémoire après les tests de rétention et/ou d'endurance.

D'autres modes de test, qui permettent l'accès direct à des cellules isolées du plan mémoire peuvent être utilisés, on distingue :

- le mode DMA (Direct Memory Access), il permet l'accès direct aux colonnes du plan mémoire en court-circuitant toute la logique d'adressage et autorise la lecture du courant traversant la cellule sélectionnée. Ce mode peut être utilisé pour extraire les caractéristiques $I_d(V_{gc})$ et $I_d(V_{ds})$ du point mémoire.
- le mode CDMA (Current Direct Memory Access), il s'agit d'un mode de lecture rapide de la mémoire par comparaison entre le courant traversant la cellule mémoire et un courant de référence.

Certaines mémoires EEPROM permettent aussi de placer les lignes et colonnes du plan mémoire dans des conditions d'utilisation critiques (ou condition de « stress »). Ce stress peut être statique ou dynamique. Ces tests sont réalisés en appliquant une haute tension sur les lignes ou les colonnes pendant un temps donné, de manière à mettre en cause les lignes ou colonnes dont les cellules mémoire présentent des oxydes défectueux.

2 Rendement des mémoires non volatiles

a. Le concept de qualité

Comme dans tous les domaines, la qualité des produits semi-conducteurs est exigée de plus en plus par le client. Aujourd'hui l'approche qualité a évolué, il ne suffit plus que le produit réponde à la description de sa fiche technique en terme de fonctionnalité. La qualité est devenue un concept global et se construit à chaque phase d'élaboration du produit (conception, fabrication, test, assemblage) pour donner un maximum de fiabilité au produit fini. Cette approche est appelée « assurance qualité ».
La phase de production des puces est un processus complexe qui compte de l'ordre de deux cent étapes pour les technologies actuelles. Pour éviter tout défaut de fabrication successible d'avoir un impact sur la qualité du produit fini, la mise en place de procédures capables de détecter les problèmes en temps réel est devenue un élément incontournable.
Ce contrôle, pour être efficace, devra permettre de détecter toute dérive de paramètres considérés critiques, pour un produit donné. Parmi ces paramètres critiques figurent les défauts de type particulaire induits lors des différentes étapes de fabrication du produit.

b. Définition du rendement

Le rendement se définit comme le rapport entre le nombre de puces bonnes à l'issu du test sur le nombre de puces potentielles que contient la plaquette de silicium. Le nombre de puces

potentielles correspondant au nombre de puces entières, inscrites dans la surface de la plaquette et pouvant donc être, a priori, fonctionnelles.
La fabrication de circuits intégrés est très complexe et inclut une succession d'étapes qui peuvent chacune générer des pertes, affectant le rendement final. De ce fait, le rendement final est fonction de trois rendements élémentaires :
Le rendement mécanique (R_m) : il correspond au rapport entre le nombre de plaquettes en sortie de production et le nombre de plaquettes lancées en début de production. Tout au long du processus de fabrication, un certain nombre de plaquettes vont être rejetées. Différentes causes vont être à l'origine de ces rejets :

- plaquettes cassées dans les équipements ou suite à une erreur de manipulation,
- plaquettes non conformes suite à une panne d'équipement,
- erreur de recette lors d'une étape de fabrication,
- plaquettes hors spécification lors de la mesure d'un paramètre critique (épaisseur des couches, dimension des motifs, mauvais alignement d'un niveau de masque par rapport au précédent …).

Ce rendement tient compte de tous les contrôles effectués durant chaque étape de fabrication. Il est toutefois important de noter que ces contrôles sont effectués par échantillonnage et sur une partie des plaquettes seulement.

Le rendement paramétrique (R_p) : les premiers test sous pointes sont les tests électriques qui vont vérifier que les structures de base constituant le circuit intégré remplissent bien le cahier des charges du point de vue électrique (tensions de seuils, résistances…). Ce test est effectué sur la totalité des plaquettes. Le rendement paramétrique sera le rapport entre le nombre de plaquettes bonnes et le nombre total de plaquettes testées.

Le rendement au test sous pointe (R_t) : après les tests électriques, chaque puce va subir une série de tests de manière à valider sa fonctionnalité. A cette étape, on obtient pour chaque plaquette le nombre final de puces bonnes. Ceci va nous donner dans un premier temps le rapport du nombre de puces bonnes sur le nombre de puces potentielles, exprimé en %. Ce pourcentage a un rôle capital. En effet, une limite de rejet est fixée pour chaque produit et toute plaquette n'atteignant pas ce niveau requis sera éliminée, soit pour des raisons économiques, soit pour des raisons de qualité.
Les plaquettes n'atteignant pas le niveau de rendement requis étant retirées, le rendement au test sous pointe est donné par le rapport entre le nombre de puces bonnes et le nombre de puces potentielles avant le test sous pointe.

Ainsi, le rendement final R_f est donné par le produit des trois rendements intermédiaires et s'exprime en pourcentage :

$$R_f = R_m * R_p * R_t \quad \text{(I. 3)}$$

c. Facteurs limitatifs du rendement

Le rendement mécanique est le résultat de toutes les mesures qui vont être effectuées à chaque étape de fabrication. Le contrôle de chaque plaquette à chaque opération étant impossible, un échantillonnage est effectué. Cet échantillonnage doit être le plus représentatif possible afin de détecter le maximum de problèmes en temps réel et de rejeter au fur et à mesure les plaquettes défectueuses. En parallèle une analyse des causes du rejet est menée et des actions

correctives sont mises en place de manière à supprimer l'origine du problème ou limiter son effet. Cette étape de correction des éventuels problèmes aura un impact direct sur le rendement final.
Le rendement mécanique sera principalement limité par :

- les performances des équipements, chaque machine a des limites techniques et doit être suivie de façon rigoureuse notamment par des opérations de maintenance préventives,
- la procédure de travail, liée à chaque opération de fabrication et effectuée suivant des procédures spécifiques dans un cadre bien défini,
- l'environnement en salle blanche qui doit être parfaitement contrôlé,
- la robustesse des procédés, suivant laquelle toute étape de fabrication doit être répétée avec le minimum de risque d'aboutir à un résultat hors spécification.

Le test électrique permet ensuite d'éliminer les plaquettes défaillantes non détectées durant la fabrication, les paramètres limitatifs à ce niveau sont les mêmes que ceux cités précédemment, avec en plus les problèmes liés au test lui même.
La dernière étape que constitue le test sous pointe va apporter de nouvelles limitations au rendement. Les tests effectués à ce niveau ainsi que les causes de rejets seront détaillés dans la partie consacrée au flot de test des mémoires non volatiles. Cependant, l'évaluation du niveau de rendement requis, fixé pour chaque produit sera analysée dans cette partie.
Il existe différents modèles permettant le calcul des limites de rejet au test sous pointe. Un modèle couramment utilisé est celui répondant à la formule de MURPHY et SEED pour le calcul du rendement :

$$R = \frac{1}{2}\left[e^{-\sqrt{n.D_0.S}} + \left(\frac{1 - e^{-n.D_0.S}}{n.D_0.S} \right)^2 \right] \quad \text{(I. 4)}$$

La loi montre que le rendement R est inversement proportionnel à chacun des facteurs S (surface de la puce en cm^2), N (nombre de niveaux de masquage), et D_0 (nombre de défauts par cm^2 et par niveau de masquage). D_0 est un contributeur majeur dans les résultats de rendements et il est d'autant plus grand que la puce à tester est grande. La surface de la puce est définie lors de la conception du produit alors que le nombre de niveaux de masque N est imposé par le processus technologique. Le paramètre important qu'il va falloir au mieux contrôler est donc la variable D_0, qui représente le nombre de défauts par cm^2 pour chaque étape de fabrication.
Le niveau de rendement requis pour chaque produit est donné par l'équation (I.4) pour un paramètre D_0 égal à dix fois le D_0 moyen considéré pour un produit donné. Toute plaquette ayant un rendement inférieur à ce niveau requis sera rejetée.

d. Impact des différentes étapes de fabrication sur le rendement

Le processus de fabrication se déroule dans quatre ateliers principaux qui sont la diffusion, la gravure, l'implantation ionique et la photolithographie.
La diffusion : on y réalise les opérations de dépôt et de croissance des couches minces sur les plaquettes, ainsi que les étapes de diffusion des ions après implantation. Ces étapes se déroulent dans des fours qui doivent subir un nettoyage systématique pour éviter que les impuretés de surface puissent diffuser dans le silicium en raison des températures très élevées. Les impuretés qui diffusent dans le substrat sont impossibles à retirer par la suite. La qualité des nettoyages (bains ou machines) va donc avoir une importance capitale et devra bénéficier d'un suivi particulier à travers des séries de mesures.

La gravure : elle consiste à graver la couche déposée précédemment sur la plaquette (nitrure, oxyde, polysilicium…). Cette gravure peut être de deux types : « humide », dans un bain d'acide, par exemple ou « sèche », par gravure plasma.
Concernant le nettoyage des bains, les gravures humides présenteront globalement les mêmes inconvénients que ceux évoqués dans la partie consacrée à la diffusion. Les chambres de gravure plasma peuvent, elles aussi, contribuer à l'apparition de défauts pouvant nuire à la qualité du produit. Les gaz utilisés doivent obéir à des normes très strictes de pureté. Les fréquences de nettoyage des chambres et des machines sont déterminées et suivies avec la plus grande rigueur.

L'implantation ionique : elle permet entre autre, de réaliser les sources et drains des transistors ainsi que les caissons de type n et p. Les défauts générés lors de cette étape sont principalement de type particulaire. Pour les recettes à forte implantation, notamment en Arsenic, il se produit un phénomène de « Wafer Charging » : à la surface du transistor, durant l'implantation, une accumulation de charges très importante a lieu qui peut entraîner des risques de claquage de l'oxyde de grille. La solution a été d'envoyer, en même temps, un petit faisceau d'électrons qui va neutraliser les charges positives dues aux ions d'Arsenic et ainsi minimiser l'accumulation de charges sur les transistors.

La photolithographie : on y réalise toutes les géométries désirées par le dépôt de résine photosensible qui sera exposée puis développée. On obtient ainsi un masque de résine représentatif du motif à graver. Ce motif est l'image d'un masque (ou « réticule ») projeté sur la surface de la plaquette recouverte de résine.
Ici, les opérations de dépôt de résine, d'exposition UV, de développement et de durcissement de la résine seront répétées pour chaque niveau de masque. Des éventuels problèmes de masquage vont donc eux aussi se répéter. Un contrôle régulier de l'air des salles blanches ainsi que de la présence de particules dans les produits et équipements utilisés est donc primordial. Il existe un paramètre important qui est la quantité de lumière que reçoit la résine durant l'exposition. En effet, l'énergie envoyée a une conséquence directe sur la dimension des motifs que l'on veut réaliser.

3 Outils et méthodologie pour l'analyse du rendement

a. Les contrôles de défauts existants et leur limitation

Le contrôle particulaire a pour but de détecter le nombre de défauts présents à la surface d'une plaquette. C'est la mesure d'une densité de défauts que l'on peut trouver sur une plaquette par unité de surface. L'unité de mesure utilisée est le nombre de défauts par cm^2 ou par plaquette.
Un défaut particulaire est représenté par une particule générée par le procédé de fabrication ou apportée par l'environnement, pouvant entraîner un dysfonctionnement de la puce. Ce peut-être aussi une rayure ou un déficit de matériaux dans une zone donnée.
Le défaut est dit « tueur » (aussi appelé « killer defect ») quand il empêche la puce de fonctionner. On définit le « killer ratio » qui représente la probabilité qu'a un défaut de tuer la puce (i.e. puce rejetée à l'issu du test). Cette probabilité est liée à la dimension de la particule et à sa localisation sur la puce [CAR93] [WIL94].
La densité de défauts générés sur les plaquettes de silicium est contrôlée à différentes étapes de fabrication. Les provenances des contaminations sont diverses : air ambiant, étapes de diffusion, implantation, gravure, photolithographie, comportement humain…
Les contrôles des défauts sont nombreux et obéissent à des procédures particulières. Ces contrôles passent par des inspections de la surface des plaques après des étapes de fabrication

critiques, de manière à évaluer la densité de défauts. Ils ciblent aussi les équipements, les chambres de dépôts, les chambres de gravure, le liquide utilisé lors des étapes de nettoyage, la qualité de l'air…
Les étapes d'inspection se heurtent aux limites des appareils de détection de défauts. En effet, il est impossible de détecter tous les défauts, mais uniquement ceux présentant suffisamment de « relief », comme les particules. De plus la taille des particules est souvent approximative puisqu'elle dépend dans de nombreux cas uniquement de la quantité de lumière réfléchie (une particule très réfléchissante mais de petite taille sera considérée comme grande).
Cependant, une grande partie de la vaste catégorie de défauts qui ne peuvent pas être observés directement sont révélés lors du test sous pointe.

b. Outils de détection et d'analyse de défauts

Les systèmes d'inspection des plaquettes localisent, mesurent la taille, le nombre et la position des défauts détectables. Ces appareils utilisent des méthodes de mesure sans contact, car tout contact est source de contamination.
Les machines d'inspection sont constituées d'un système de chargement et de déplacement des plaquettes, d'un système d'inspection optique (et souvent de digitalisation de l'image) et d'un système de traitement d'images.
Certains systèmes d'inspection sont basés sur le principe d'analyse et de comparaison d'images. Une source de lumière éclaire le substrat sur une surface de petite dimension pour ensuite venir balayer toute la plaquette. Un système de détection, associé à une caméra, analyse le signal reçu à travers un ensemble optique [HA92] [FAN98]. Ce système d'analyse est embarqué dans des appareils d'inspection de type « KLA-Tencor ».

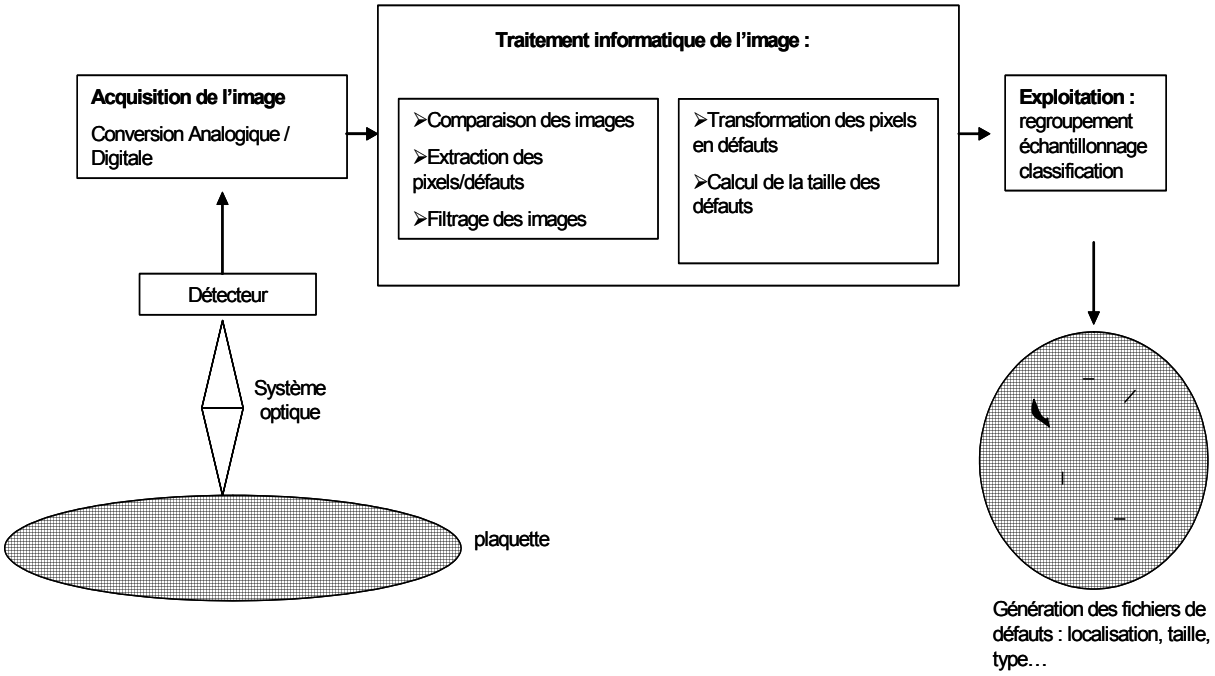


Figure I. 26 Système d'inspection de défauts.

Une méthodologie d'analyse de défauts comprend trois étapes principales : l'acquisition de l'image, le traitement informatique de l'image qui va permettre la classification des défauts et l'exploitation des résultats d'analyse (figure I.26).

L'acquisition de l'image : l'image optique est transmise vers un détecteur qui va analyser l'image reçue. L'image est divisée en plusieurs parties élémentaires et chacune d'elle va émettre un rayonnement différent. Le détecteur, qu'on peut assimiler à une grille constituée de plusieurs éléments de taille identique (pixels) va mesurer cette intensité lumineuse en chacun de ces points et va la coder sur 8 bits (généralement 256 niveaux de gris pour une image noir et blanc).
Lors de l'inspection, la plaquette est généralement posée sur une table mobile qui se déplace de façon continue de manière à être balayée par le détecteur fixe.
La détection de défauts se fait par comparaison d'images. Pour l'inspection des circuits mémoire, il y a comparaison des cellules adjacentes et identiques entre elles. Pour les parties logiques la comparaison se fera par comparaison d'une puce adjacente à une autre. Les défauts sont donc détectés par comparaison d'une image à une image de référence. La précision de la détection est assurée par un système de grossissement de l'image qui fixe la résolution du détecteur.

Le traitement informatique de l'image : pour une meilleure compréhension des défauts et de leurs origines, la taille de ces défauts est une indication primordiale. En effet, en plus du type et du nombre de défauts, les informations concernant leur taille vont nous permettre de déterminer si le défaut est « tueur » pour le circuit.
La détermination de la taille des défauts est délicate puisque ces derniers sont traduits en pixels. De ce fait, tout défaut dont la taille est inférieure à la taille du pixel sera comptabilisé comme ayant la taille d'un pixel. Cette information est complétée par la luminosité du pixel puisque seuls les pixels ayant un niveau de gris supérieur à un seuil défini seront comptabilisés. Les limites de cette méthode de classification concernent le manque d'informations concernant le taux de remplissage (surface réelle du défaut), puisque ce dernier est décrit par sa largeur et sa hauteur. L'approximation la plus utilisée calcule la taille du défaut en utilisant la minimum des trois valeurs suivantes : largeur, hauteur et racine carrée de l'aire du défaut (exprimée en pixels).

L'exploitation des résultats d'analyse : à l'issue de toutes les étapes d'inspection, plusieurs informations sont disponibles comme :

- la répartition géographique des défauts sur la plaquette (figure I.27),
- les résultats chiffrés de l'inspection : plaquette, nombre de défauts, densité…,
- la répartition du nombre de défauts par catégorie,
- la répartition des défauts par taille,
- la photographie de certains défauts.

Certains appareils utilisent d'autres principes de mesure pour déterminer et détecter la position et le nombre de défauts présents sur une plaquette de silicium. Dans le cas des appareils de la famille « Tencor surfscan » [WIL94], la lumière est focalisée sur la surface du substrat et un rayon est réfléchi suivant le même angle que le rayon incident. Lorsqu'une contamination se trouve à la surface du substrat, une petite partie du rayon incident sera réfléchie suivant un autre angle. En plaçant le détecteur à un endroit approprié, la lumière dispersée pourra être détectée. Le signal détecté est ensuite transmis à un photomultiplicateur qui va le transformer en signal électrique. La forme du signal obtenu par le photomultiplicateur est une gaussienne. L'amplitude, mais aussi la forme de la gaussienne apportent les informations relatives à la particule. La taille exacte de la particule est obtenue à partir de courbes de calibration réalisées à partir de billes en latex dont le diamètre est connu. Un traitement informatique permet ensuite une localisation précise de la particule.

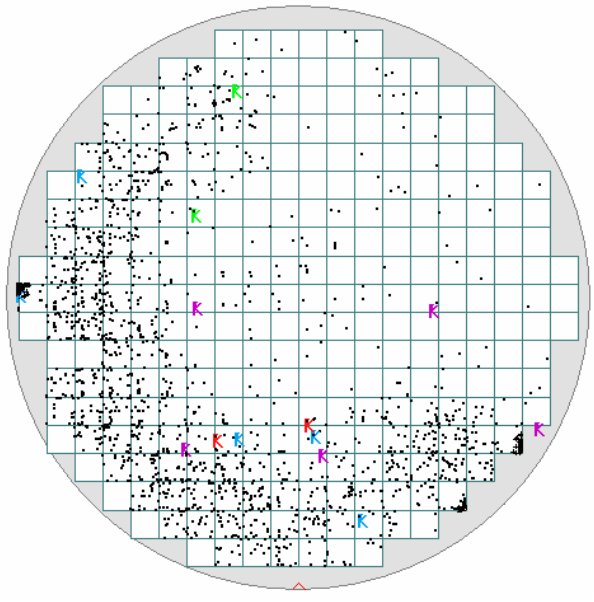


Figure I. 27 Répartition géographique des défauts.

4 Le flot de test des mémoires non volatiles

a. Classes de rejet

Les programmes de test appliqués aux mémoires non volatiles sont composés de séquences de test spécifiques. Les puces qui fournissent des réponses erronées en réponse aux signaux appliqués durant un test déterminé sont associées à une classe de rejet ou « binning ».
Dans le cas des mémoires non volatiles de type EEPROM, les classes de rejets varient entre zéro et quinze. Les puces mémoires ne présentant aucune défaillance sont associées à la classe de rejet numéro zéro. Les puces qui nécessitent l'utilisation de la redondance pour éviter le rejet lors de certains tests fonctionnels sont associées à une classe de rejet numérotée un. Ces puces sont considérées comme des puces bonnes.
Les différentes broches d'un circuit mémoire assurent principalement quatre types de fonctions : entrées, sorties, contrôles et alimentations ou masses. A travers une séquence précise d'impulsions appliquées sur les broches d'entrées-sorties et de contrôle, l'utilisateur de la puce peut effectuer un nombre limité d'opérations sur le dispositif. Chaque opération est généralement composée d'une suite d'opérations élémentaires (effacement, écriture et lecture). Ce mode d'accès est spécifique à l'utilisateur du dispositif, il est appelé mode d'accès client (ou « user mode »).
Un second mode d'accès au dispositif est le mode test. Il permet l'exécution de toutes les opérations élémentaires de manière indépendante. Des modes d'analyse plus complexes, comme le mode « DMA », « CDMA » et l'utilisation de méthodes de programmation particulières sont disponibles durant le mode test.
La figure I.28 représente le flot de test des mémoires non volatiles. Ce flot de test s'applique aux mémoires non embarquées (dans des systèmes comme les cartes à puces, le flot de test est beaucoup plus dense). Chaque test élémentaire sera décrit dans les parties suivantes.

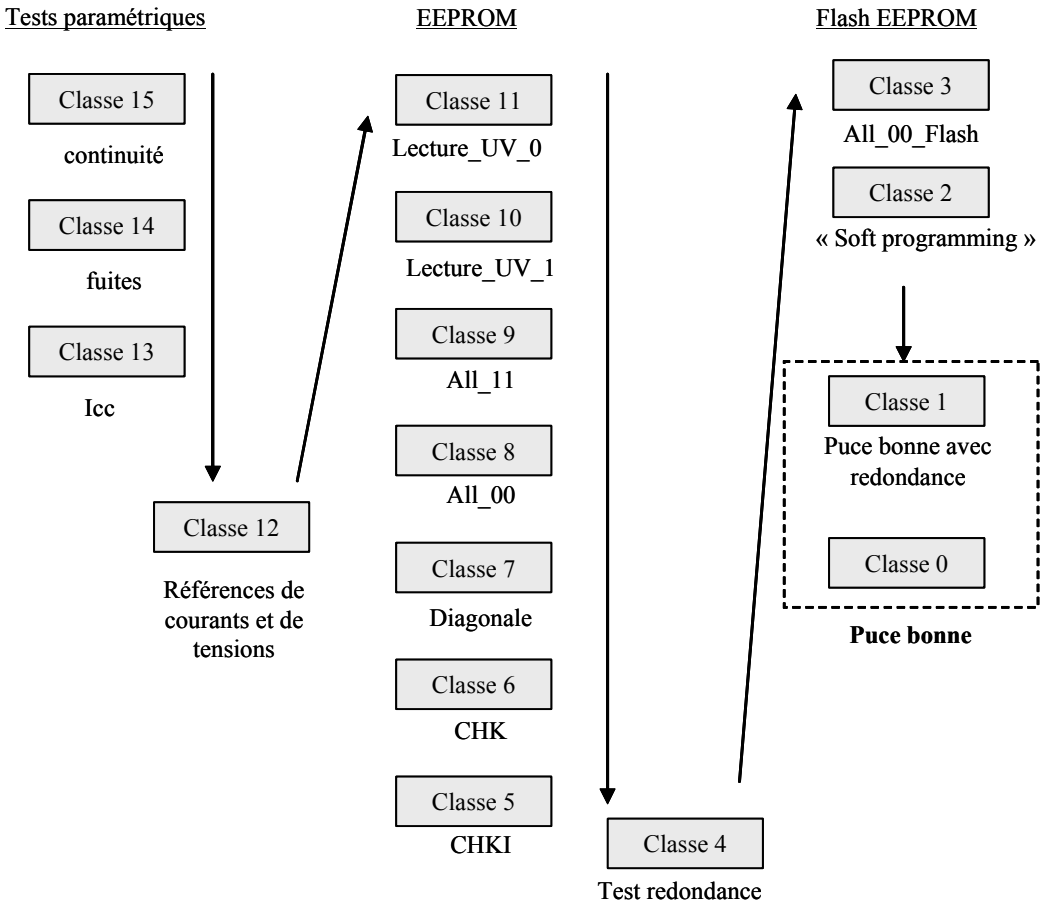


Figure I. 28 Flot de test des mémoires non volatiles.

b. Tests paramétriques

Ces tests ont pour but de contrôler l'intégrité de base du circuit. Les tests effectués à ce niveau détectent d'éventuels dysfonctionnements des broches du composant ainsi qu'une consommation anormale.
Les classes de rejets 15 et 14 contrôlent l'intégrité des broches du circuit et la nature du contact entre les broches et la carte à pointe. La classe 13 détecte les anomalies de consommation au niveau de la broche d'alimentation du dispositif.
La classe 12 permet de vérifier que la tension de lecture V_{ref}, correspond à un courant de lecture donné I_{ref}, sur la caractéristique $I_d(V_{gc})$ entre dans les spécifications. Ce niveau de tension est déterminé à partir d'une cellule de référence dans l'état électrique vierge. La tension de lecture est calculée en faisant varier la tension de grille V_{gc} jusqu'à ce que le courant traversant la cellule vierge soit égal à une valeur de référence notée I_{ref}. I_{ref} peut être fixé à 5 µA ou 10 µA suivant la technologie. La tension de lecture déterminée est capitale puisqu'elle sera fournie par un circuit référence de tension pour toutes les phases de lecture du flot de test.
Un rejet de classe 12 correspond à une valeur de la tension de lecture en dehors des spécifications. Dans l'exemple de la figure I.29, la tension de lecture $V_{gc\_5\mu A}$ est déterminée à partir d'un courant I_{ref} fixé à 5µA. Et cela, à partir la caractéristique $I_d(V_{gc})$ de la cellule de référence dans son état électrique vierge.

Au niveau du programme de test, des algorithmes simples (dichotomie) permettent de détecter la tension de lecture par variation de la tension de grille. Lorsqu'un courant de valeur I_{ref} traversant la cellule est détecté, la valeur de la tension de grille correspondante est mémorisée.

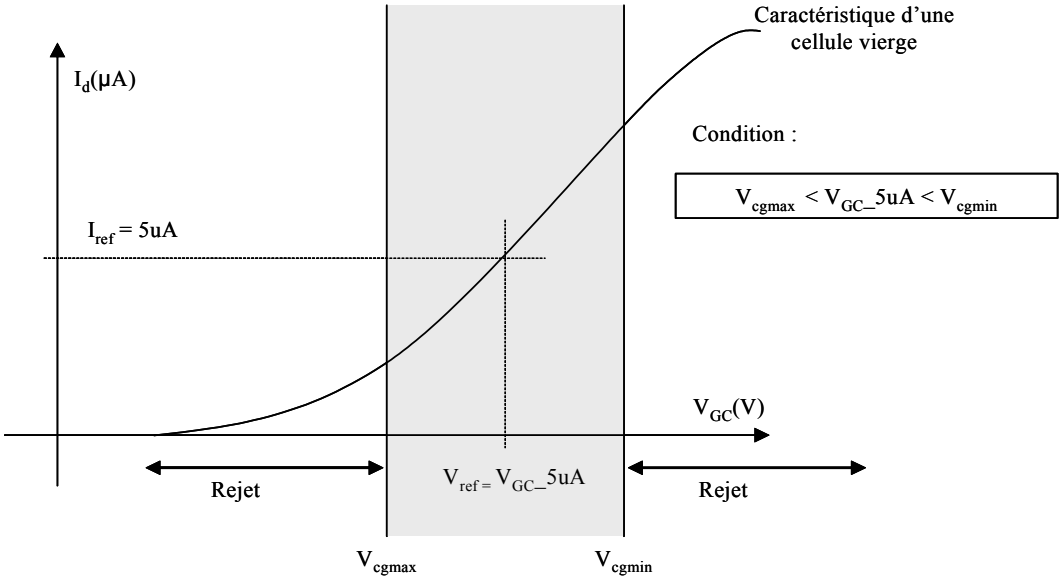


Figure I. 29 Détermination de la tension de lecture.

c. Tests fonctionnels

Les premiers tests fonctionnels sont appelés tests de « virginité ». Ils permettent de détecter les cellules vierges qui ne fournissent pas assez, ou trop de courant pour des tensions de lecture de grille de contrôle données.
La figure I.30a représente le cas d'une lecture de cellules vierges à l'état logique 1 (« UV_11 »). Lors de ce type de test, toutes les cellules qui fournissent un courant inférieur au courant de lecture I_{ref}, pour la tension de grille V_{GC_1} donnée sont défectueuses. Ces cellules défectueuses sont lues à l'état logique 0 (i.e. leur tension de seuil V_{GC_d} est trop élevée) et la puce est rejetée en classe 10. La figure I.30b représente la distribution des cellules d'un plan mémoire et met en évidence une cellule défectueuse dont le seuil de conduction à un courant de référence donné est obtenu pour une valeur de la tension de la grille de contrôle V_{GC_d} trop élevée.

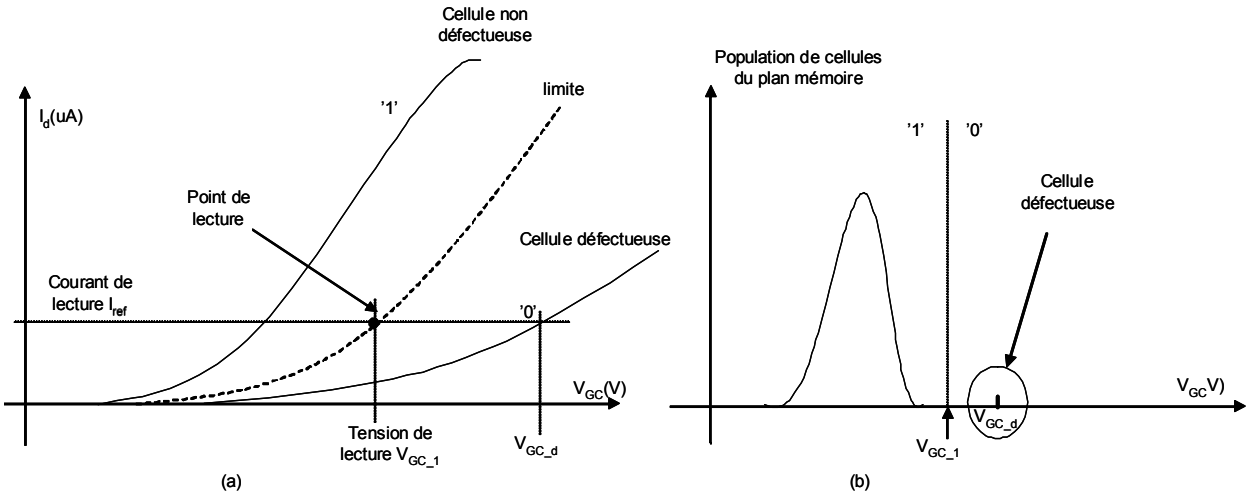


Figure I. 30 Lecture d'une cellule vierge à l'état logique 1.

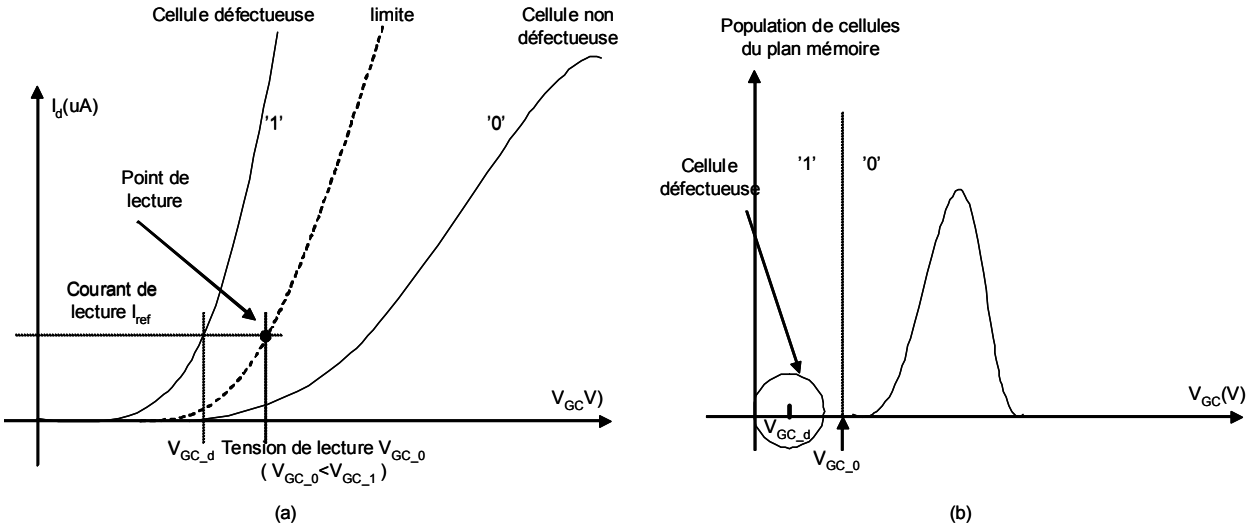


Figure I. 31 Lecture d'une cellule vierge à l'état logique 0.

La figure I.31a représente le cas d'une lecture de cellule vierge à l'état logique 0 (« UV_00 »). Toutes les cellules qui fournissent un courant supérieur au courant de lecture pour la tension de grille V_{GC_0} donnée sont défectueuses. Dans ce cas, ces cellules défectueuses sont lues à l'état logique 1 et la puce est rejetée en classe 11. La figure I.31b met en évidence la présence d'une cellule défaillante dont le seuil de conduction est obtenu pour de faibles valeurs V_{GC_d} de tension de la grille de contrôle.
Ces tests permettent d'assurer que toutes les caractéristiques vierges des cellules du plan mémoire appartiennent à une fenêtre de spécification définie par V_{GC_0} et V_{GC_1}. On associe à ce type de rejets des défauts liés à des contacts résistifs, des discontinuités au niveau des lignes de métal ou des particules résiduelles engendrées par des attaques chimiques lors de gravures. Durant ces tests, la tension de drain est fixée par le circuit de lecture à environ 0.6V.

Le deuxième type de tests fonctionnels va permettre de vérifier la fonctionnalité de la mémoire, c'est-à-dire la capacité de mémorisation des cellules. Ces tests se décomposent en deux étapes : une première étape de programmation va imposer à chaque cellule du plan mémoire un état logique donné (0 ou 1 comme le montre la figure I.32). Ensuite, une deuxième étape de lecture du plan mémoire va vérifier l'intégrité des informations écrites dans la mémoire.
Le test « All_11 » permet d'écrire toutes les cellules du plan mémoire en les portant à un état logique 1. Cette écriture est réalisée durant une phase de programmation qui met en jeu la pompe de charge. Une lecture est ensuite effectuée en adressant chaque cellule du plan mémoire. Cette lecture est réalisée à la tension de grille $V_{GC\_lecture}$ déterminée précédemment.
Lors de la phase de lecture, si des cellules mémoires mal ou pas écrites fournissent la valeur logique 0 au niveau de l'amplificateur de lecture, la puce est rejetée en classe 9.
De la même manière les puces rejetées en classe 8 sont des puces qui contiennent des cellules mémoires mal ou sous effacées.
Il est à noter que dans le cas des mémoires EEPROM, chaque phase de programmation est composée d'une phase d'effacement suivie d'une phase d'écriture, le signal de programmation double rampe étant généré à partir d'un organe interne au produit EEPROM qui est la pompe de charge.

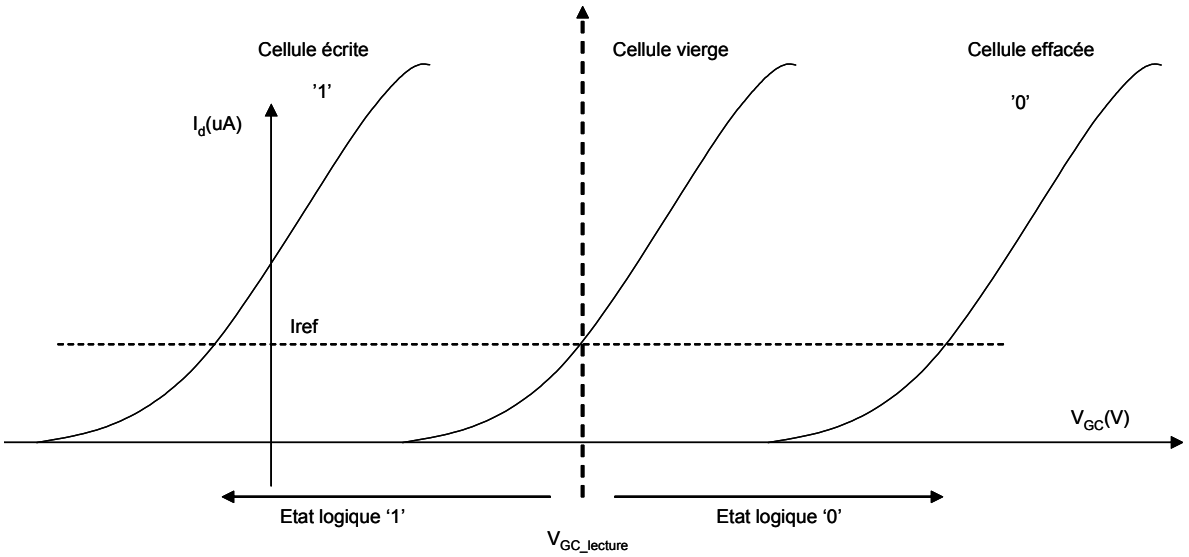


Figure I. 32 Lecture et états logiques (mémoires EEPROM).

Les tests qui suivent les étapes de programmation vont consister à écrire des motifs particuliers sur tout le plan mémoire.
Les figures I.33a et I.33b présentent un test qui consiste à écrire une diagonale dans le plan mémoire. Ce type de test se décompose en une phase de programmation, où toutes les cellules formant la diagonale du plan mémoire vont être écrites. Ensuite, une phase de lecture va relire tout le plan mémoire de manière à évaluer le nombre de cellules écrites. Ce nombre est comparé au nombre de cellules qui composent la diagonale. Un rejet à ce type de test survient lorsque le nombre de cellules de la diagonale est différent du nombre de cellules lues comme étant écrites.
Le test qui consiste à écrire une diagonale cible tous les problèmes liés au décodage des adresses. Il peut aussi mettre en évidence d'éventuels courts-circuits au niveau des colonnes et des lignes du plan mémoire. Cette classe de rejet, numérotée 7, est difficile à analyser puisqu'elle peut aussi mettre en cause un dysfonctionnement au niveau matériel (niveaux de tensions appliqués, carte à pointes…) ou provenir d'une défaillance des circuits internes à la pompe de charge qui rend la programmation difficile.

1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1

(a)

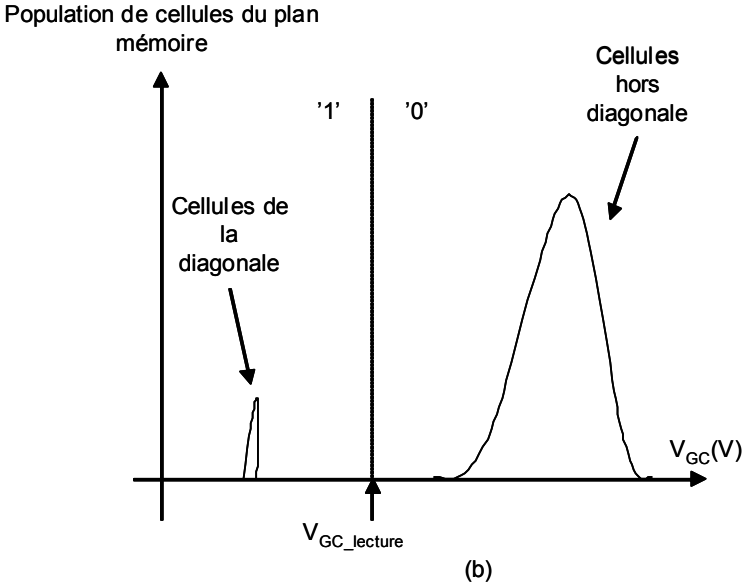


Figure I. 33 Le test « diagonale »

Le test qui suit l'écriture et la lecture de la diagonale va consister à programmer tout le plan mémoire de manière à réaliser un « damier » constitué de cellules écrites et de cellules effacées (CHK). Ce test est suivi d'un test appelé « damier inverse » (CHKI). Ces tests consistent à écrire des données complémentaires dans les cellules topologiquement adjacentes. Une défaillance typique révélée lors de ce test est un court-circuit au niveau du polysilicium constituant la grille flottante des cellules mémoires. Les figures I.34a et I.34b illustrent ce type de test.

1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1

(a)

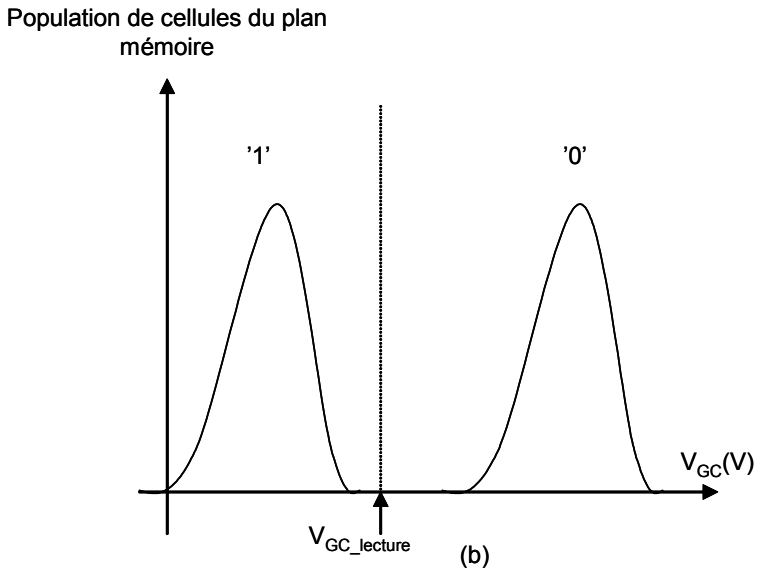


Figure I. 34 Le test « damier »

Durant le flot de test et en particulier lors de l'exécution des tests fonctionnels, l'adresse de toutes les cellules défaillantes est mémorisée au sein d'une mémoire intermédiaire, disponible sur les équipements de test (souvent appelée mémoire « cache »). Les cellules défaillantes sont ainsi comptabilisées. Suivant la redondance mise en place sur la puce, les cellules défaillantes peuvent être corrigées par substitution. Si le nombre de cellules défaillantes est trop important et ne peut être pris en charge par les lignes et colonnes de redondance, un rejet fonctionnel survient pour le test considéré.
Dans le cas favorable d'une redondance suffisante, un ré-adressage automatique des signaux de lecture et de programmation est effectué pour les lignes et colonnes défaillantes, en direction des lignes et colonnes de redondance. L'adresse des lignes et colonnes défaillantes est inscrite dans une zone mémoire distincte à l'extérieur du plan mémoire (cf. §B.3). Lorsque des problèmes d'accès ou de programmation de cette mémoire externe sont détectés, la puce est rejetée en classe 4.

La classe de rejet 3 concerne particulièrement l'effacement sectoriel des mémoires flash EEPROM. Les temps d'effacement et la forme des distributions des cellules effacées (qui mettent en évidence des cellules sur-effacées ou mal effacées) sont très sensibles au contexte technologique, et aux polarisations appliquées lors des phases de test. Les cellules sur-effacées dont les tensions de seuil sont négatives sont prises en charge par la redondance.
La classe de rejet 2 est liée aux cellules possédant des tensions de seuil trop basses situées sous un niveau d'effacement fixé par des cellules de référence. La tension de seuil de ces cellules est ramenée dans les spécifications par le biais d'une programmation spécifique appelée « soft programming ». Si le temps de programmation associé à cette phase de test dépasse une certaine limite de temps donnée par les spécifications (trop de cellules à programmer), la puce est rejetée en classe 2.

D'autres tests, appliqués le plus souvent aux mémoires non volatiles permettent de mettre en évidence les défaillances affectant l'oxyde inter polysilicium (ONO) et l'oxyde tunnel. Il s'agit respectivement des tests de stress des électrodes de grille (« gate stress ») et des tests de stress des électrodes de drain (« drain stress »).

d. Tests de rétention et d'endurance des mémoires EEPROM

Le premier flot de tests est suivi d'un deuxième flot de tests qui va vérifier la capacité des cellules mémoire à sauvegarder l'information. Le plan mémoire est en général dans un état de départ effacé (grille flottante chargée en électrons). Ensuite, les puces sont soumises à de hautes températures, en général 250°C pendant 24 heures, de manière à simuler un vieillissement du dispositif et déceler des fuites de charges à travers l'oxyde tunnel, l'ONO ou les aspérités du polysilicium [SOB95] [KUM01].
Cette perte de charges est évaluée à partir du décalage de la distribution des cellules électriquement effacées du plan mémoire.
A la fin du premier flot de test, toutes les cellules du plan mémoire sont effacées et la valeur de la tension de seuil la plus basse des cellules effacées est mémorisée dans une partie dédiée de la mémoire.
Après avoir soumis les puces à des températures élevées, une nouvelle lecture avec une tension de grille V_{GC} variable est effectuée de manière à extraire la distribution des cellules après le test en rétention. Durant cette phase de lecture, on détermine la tension de seuil la plus basse des cellules du plan mémoire.
Le décalage des tensions de seuil ΔV_T est donné par la différence de la tension de seuil la plus faible avant rétention et la tension de seuil la plus basse après rétention comme présenté figure I.35. Ce décalage doit être contrôlé et maintenu dans un intervalle fixé par les spécifications (400 mV pour les mémoires EEPROM).

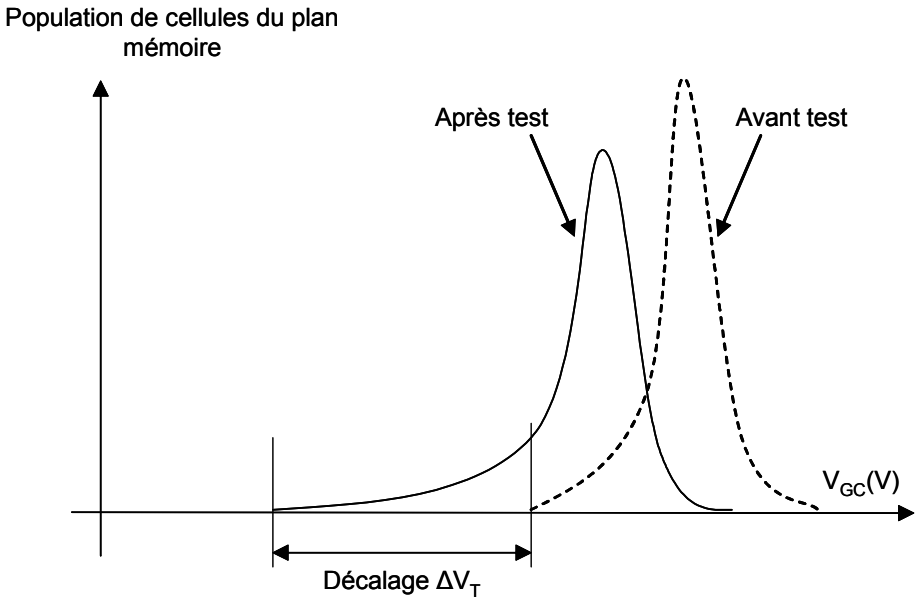


Figure I. 35 Le test en rétention.

Les tests d'endurance (aussi appelés tests de « cyclage ») consistent à effectuer plusieurs cycles d'effacement/écriture sur les cellules du plan mémoire. L'endurance d'une mémoire non volatile est principalement limitée par la tension de claquage de l'oxyde tunnel.

Ce claquage d'oxyde étant généralement provoqué par des points faibles (défauts) situés dans l'oxyde.
Les tests d'endurance sont souvent associés aux tests de rétention suivant un ordre précis : les cellules du plan mémoire subissent dans une première phase des tests d'endurance (100Kcycles). Puis, les puces sont soumises à un test de rétention. La figure I.36 est une superposition des distributions des cellules du plan mémoire dans son état initial effacé, après rétention et après les tests d'endurance. La forme écrasée de la distribution après le test d'endurance et le résultat de la fragilisation des oxydes qui entourent la grille flottante.

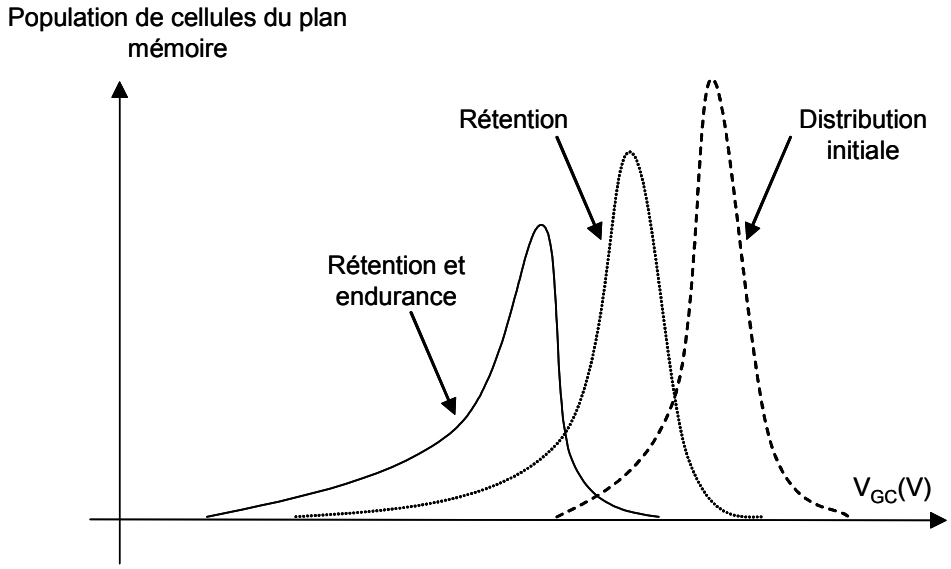


Figure I. 36 Les tests en rétention et en endurance.

e. « Bitmap » et analyse de redondance dans le plan mémoire

L'analyse « bitmap » cible les défaillances qui affectent le plan mémoire. Le « bitmap » est une représentation topologique des cellules défaillantes d'un plan mémoire. Il est obtenu après que tous les tests fonctionnels aient été appliqués au circuit mémoire. La localisation des cellules défaillantes fournit des signatures électriques (disposition particulière de cellules défaillantes) à partir desquelles il est possible dans certains cas, de remonter à l'origine physique de la défaillance.
Une observation visuelle est effectuée par la suite pour localiser le défaut. Il est souvent nécessaire de retirer des couches de matériaux successives (gravure plasma, bains d'acide) constituant le circuit ou de procéder à des coupes pour remonter à la nature précise du défaut. Ce défaut est ensuite analysé de manière à situer le problème de manière précise dans les étapes de fabrication du produit. Les modifications appropriées sont alors apportées de manière à améliorer le rendement.
Pour les contaminations provenant de particules de petite taille, qui affectent une seule ligne du plan mémoire par exemple, il est possible de substituer la ligne contenant le défaut par une ligne de « rechange » généralement située sur les côtés du plan mémoire. Ces lignes (ou colonnes) de substitution reçoivent les adresses des lignes (ou colonnes) défectueuses.
Le concept de redondance pour l'amélioration du rendement appliqué aux mémoires est connu depuis longtemps. En 1967, TAMARU et ANGELL [TAM67] proposent une nouvelle technique de remplacement des éléments défaillants d'une structure répétitive après avoir localisé l'élément défaillant lors de procédures de test spécifiques. Plus précisément, cette technique a été utilisée pour des circuits mémoires LSI (Large Scale Integrated circuit). Il a

été proposé l'utilisation de lignes de réparation dans le but d'augmenter significativement le rendement puisque seules les lignes non défaillantes de la mémoire étaient connectées au boîtier lors de la phase d'assemblage. En 1969, CHEN élargit l'utilisation de cette technique en utilisant à la fois des lignes et des colonnes de redondance [CHE69].
Sur certains produits actuels, il est possible de substituer des rangées de lignes ou de colonnes. Ce nombre de colonnes ou lignes de redondance est limité par les contraintes de surfaces générées par les éléments de redondance ainsi que par leur gestion. Il est à noter qu'il n'est prévu aucune redondance pour les circuits logiques de la mémoire.

CHAPITRE II
Diagnostic de Défauts Géométriques dans la Cellule Mémoire EEPROM

I. La structure MOS dans les MNV

A. Modélisation du transistor MOS

1 Rappels sur la structure MOS

a. Régimes de fonctionnement du transistor MOS

La structure basique d'un transistor MOSFET de type n est reportée figure II.1. Comme son nom l'indique (Metal-Oxide-Semiconductor Field Effect Transistor), le transistor MOSFET ou plus simplement MOS est constitué d'un substrat semiconducteur faiblement dopé de type p (noté p) sur lequel a été déposé une fine couche isolante (S_iO_2) d'épaisseur T_{ox}. Une couche conductrice (en métal ou en polysilicium), formant l'électrode de grille est ensuite déposée sur cette couche isolante. De part et d'autre de la grille, deux régions fortement dopées de type n (notées n^+), de profondeur X_j, forment les électrodes de source et de drain du transistor. En raison du procédé de fabrication, la grille recouvre légèrement les régions de source et de drain. La zone de substrat, sous la grille, située entre les deux zones de diffusion de drain et de source, représente le canal du transistor. On distingue la longueur effective du canal L_{eff}, formée en excluant l'empiétement des régions de source et de drain sous la grille, de la longueur de grille L_g. Les principaux paramètres de cette structure sont L_{eff}, T_{ox}, la largeur du transistor W et le dopage du substrat de type p, N_a.
Les potentiels appliqués sur les électrodes de grille, de drain, de source et de substrat (bulk) sont respectivement V_g, V_d, V_s et V_b.

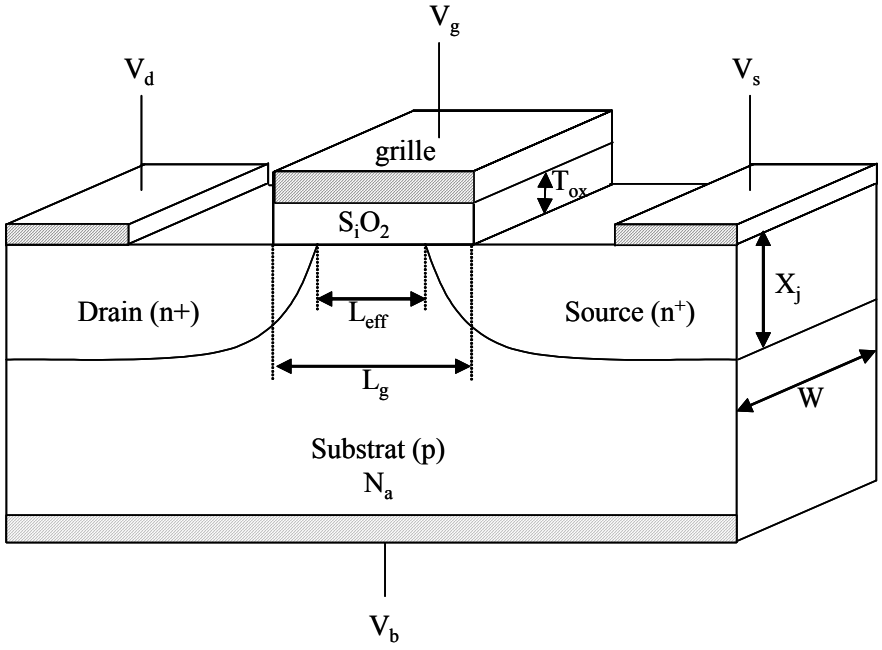


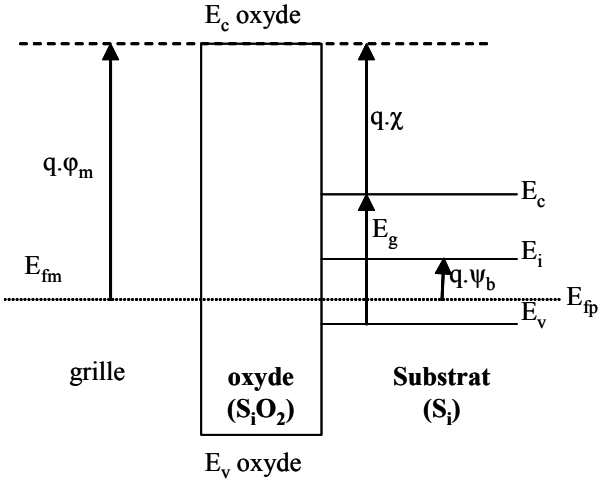
Figure II. 1 Structure basique d'un transistor MOS de type n.

Le principe de fonctionnement du MOS peut être schématisé à partir des diagrammes de bande d'énergie, représentatifs de ses différents modes de fonctionnement. Pour l'instant, nous supposons que les électrodes de source et de drain du transistor sont à la masse, le potentiel de référence étant le potentiel de substrat ($V_{sb} = V_{db} = 0$).

Dans ces conditions, lorsqu'une tension V_{gb} est appliquée entre la grille et le substrat, la structure initiale des bandes d'énergie du transistor MOS schématisée figure II.2 est modifiée à l'interface S_i-S_iO_2. Le diagramme d'énergie de la figure II.2 est relatif à une structure MOS sans défaut, dont les particularités sont les suivantes [BOU99] :

- il n'y a pas de différence entre les travaux de sortie du métal et du semiconducteur,
- il n'y a pas de charges électriques dans la couche d'oxyde,
- il n'y a pas d'états électroniques à l'interface oxyde-semiconducteur.

Il en résulte que dans le diagramme d'énergie de la structure MOS idéale, les bandes du semiconducteur sont plates en l'absence de toute polarisation.



ψb : potentiel de substrat ($q.\psi b = E_i - E_{fp}$)

E_{fp} : potentiel électrochimique (ou niveau de Fermi) du semiconducteur de type p

E_{fm} : potentiel électrochimique du métal

$q.\chi$: Affinité électronique du semi-conducteur

$q.\Phi m$: Travail de sortie (grille métallique)

Figure II. 2 Diagramme d'énergie de la structure MOS idéale non polarisée.

Si nous appliquons une différence de potentiel V_{gb} entre la grille (considérée comme métallique) et le substrat de type p alors, l'absence de migration de charges à travers la couche d'oxyde entraîne une courbure des bandes d'énergie dans le semiconducteur. Cette courbure des bandes énergétiques est le résultat d'une accumulation de charges (électrons ou trous) au voisinage de l'interface oxyde-semiconducteur.
En fonction de la variation de la tension de grille V_{gb}, la figure II.3 met en évidence trois régimes de fonctionnement : accumulation, déplétion et inversion.

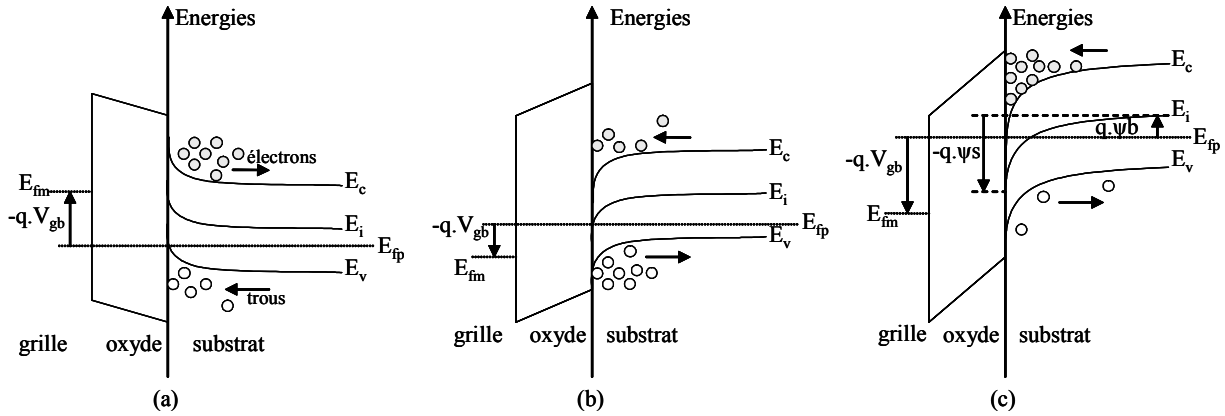


Figure II. 3 Courbure des bandes énergétiques en accumulation (a), déplétion (b) et inversion (b).

Lorsque la tension V_{gb} est négative, les potentiels électrochimiques du métal et du semiconducteur E_{fm} et E_{fp} sont décalés. Il y a accumulation de trous à la surface du semiconducteur provoquant une courbure des bandes d'énergie. On peut noter qu'à l'interface, le niveau d'énergie de la bande de valence du semiconducteur E_v se rapproche du pseudo-niveau de Fermi E_{fp}, ce qui entraîne une concentration de charges positives à la surface du semiconducteur (figure II.3a).

Lorsque la tension V_{gb} devient positive, les trous sont repoussés vers l'intérieur du semiconducteur alors que les électrons sont attirés vers la surface. A l'interface et pour ces conditions de polarisation, le niveau de Fermi intrinsèque du semiconducteur E_i se rapproche du pseudo niveau de Fermi E_{fp}. Cela entraîne une concentration de charges négatives à la surface du semiconducteur (figure II.3b).

Lorsque la tension V_{gb} continue d'augmenter, le niveau de Fermi intrinsèque E_i se rapproche encore plus du pseudo niveau de Fermi E_{fp} en surface, jusqu'à passer en dessous de ce dernier (figure II.3c). Les valeurs des concentrations d'électrons minoritaires et de trous majoritaires sont fonction des positions relatives de ces deux derniers niveaux E_i et E_{fp}. Pour de telles conditions de polarisation, nous pouvons distinguer deux régimes de fonctionnement correspondant à la faible et la forte inversion :

- l'inversion faible vérifiant la condition : $2\Psi_b > \Psi_s > \Psi_b$
- l'inversion forte étant obtenue pour : $\Psi_s > 2\Psi_b$

C'est seulement durant le régime d'inversion forte que la concentration des électrons en surface devient importante; on dit alors que le type du semiconducteur est inversé en surface par rapport au volume.

A partir de cette définition, la tension de seuil de la structure MOS, notée V_T, peut être définie comme la valeur particulière de la tension V_{gb} pour laquelle le régime d'inversion forte est atteint.

Dans le cas du transistor MOS soumis à une polarisation de drain positive ($V_{ds} \neq 0$) et pour une tension de grille V_g supérieure à la tension de seuil du transistor V_T, le phénomène suivant se produit : sous l'effet du champ électrique longitudinal créé par une différence de potentiel entre le drain et la source, un courant de porteurs minoritaires apparaît au niveau de la zone d'inversion. Ce courant sera à la fois fonction des tensions V_g et V_d puisque, d'une part, la tension V_g va moduler la conductance du canal et que, d'autre part, le champ électrique longitudinal est fonction de la tension de drain V_d.

b. Expression de la charge dans le semiconducteur

L'expression de la charge mise en jeu dans le semiconducteur à l'interface est fonction du potentiel de surface Ψ_S et peut être évaluée par l'expression suivante [MAS00] :

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

Avec un signe positif si $\Psi_s < 0$ et un signe négatif si $\Psi_s > 0$. Dans cette expression, Φ_c représente l'écart entre les quasi-niveaux de Fermi, β le potentiel thermique et n_0 et p_0 la concentration des électrons et des trous libres dans le semiconducteur. Il est important de rappeler que cette expression n'est valable que dans le cas d'un semiconducteur de type p. La figure II.4 illustre les variations de la valeur absolue de la charge dans le semiconducteur $|Q_{SC}|$

en fonction du potentiel de surface pour différentes valeurs du dopage (supposé constant) du semiconducteur.
Le potentiel de surface est une grandeur fondamentale utilisée pour décrire les caractéristiques électriques du transistor MOS dans de nombreux modèles. Il représente l'importance de la courbure des bandes en termes de potentiel, c'est-à-dire le potentiel électrostatique à l'interface oxyde de grille-substrat par rapport à la zone neutre du substrat.

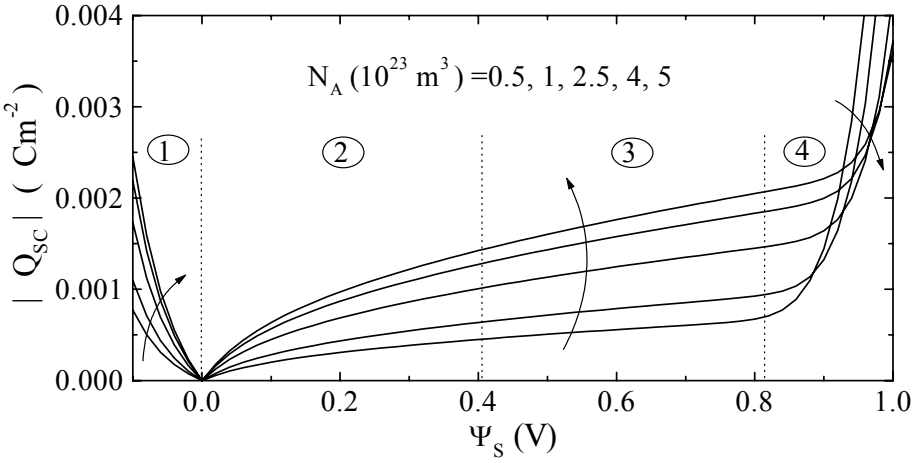


Figure II. 4 Charge dans le semiconducteur pour différents dopages substrat.

On retrouve, à partir de la figure II.4, les quatre régimes de fonctionnement du transistor :

- $\Psi_s < 0$: régime d'accumulation. Les trous, porteurs majoritaires du substrat, sont accumulés à l'interface grille-substrat.
- $0 < \Psi_s < \Psi_b$: régime de déplétion ou de désertion. Les trous sont encore présents à l'interface mais moins nombreux que dans le volume du substrat.
- $\Psi_b < \Psi_s < 2\Psi_b$: régime d'inversion. La concentration en électrons libres à l'interface devient plus importante que celle des trous.
- $\Psi_s > 2\Psi_b$: régime d'inversion forte. Les électrons libres, porteurs minoritaires du volume du semiconducteur, sont accumulés à l'interface où ils forment la couche d'inversion.

2 Modélisation du transistor MOS

Les modèles de transistor décrivent les comportements d'un composant en terme de caractéristiques électriques, principalement courant-tension (I-V) et capacité-tension (C-V), ainsi que le principe de déplacement des porteurs dans le dispositif [ARO99]. Ces modèles reflètent le comportement du composant dans toutes ses zones de fonctionnement. Dans le cas du transistor MOS, le modèle doit être valide aussi bien dans son mode de fonctionnement en accumulation qu'en inversion. Deux principales catégories de modèles existent : les modèles physiques et les modèles compacts destinés à la simulation de circuits.

a. Modèles physiques

Les modèles physiques des dispositifs sont basés sur une définition rigoureuse des paramètres géométriques, des profils de dopage, des équations de transport des porteurs et des caractéristiques des matériaux utilisés. Ces modèles sont utilisés pour prédire les caractéristiques électriques du transistor MOS ainsi que les différents phénomènes liés au transport des porteurs. Leur utilisation nécessite une forte puissance de calcul, associée à des

algorithmes de calcul performants en raison de la résolution complexe des équations couplées de la physique et du semiconducteur. L'intérêt de ces modèles est qu'ils fournissent une compréhension détaillée de l'aspect physique du fonctionnement du transistor, et qu'ils ont une réelle capacité à prédire les caractéristiques électriques du dispositif futur. Le principal inconvénient de ces modèles est leur incapacité à être utilisés dans la simulation de circuits intégrés car ils nécessitent un temps de calcul beaucoup trop élevé.

b. Modèles « compacts »

Du fait de la nature 2D, voire 3D des effets physiques qui régissent le fonctionnement des transistors MOS, il est difficile d'obtenir une formulation analytique qui soit valide dans tous les régimes de fonctionnement du transistor. Cependant, il est possible d'obtenir des formulations analytiques, basées sur la physique du composant, mais généralement limitées à un domaine de fonctionnement particulier. De tels modèles sont utilisés pour la simulation de circuits en raison de leur rapidité d'exécution.
Cette famille de modèles, appelée « modèles compacts », est utilisée dans les simulateurs de circuits standard et contient d'une part des expressions basées sur la physique et d'autre part un certain degré d'empirisme. Ces modèles peuvent être adaptés à différentes technologies CMOS à l'aide d'un certain nombre de paramètres, dans le but de décrire correctement les caractéristiques électriques du composant. Afin de rendre pratique l'utilisation du modèle, ce dernier doit être complété par des méthodes d'extraction de paramètres (ces paramètres constituent la « carte modèle » du composant).
Le choix d'un modèle de transistor est très difficile, il dépend tout autant de la précision du modèle que du domaine d'application. Il est donc nécessaire pour le concepteur de pouvoir disposer dans une technologie donnée, de plusieurs modèles fiables. Toutefois, le choix d'un modèle de transistor MOSFET doit être guidé par la prise en compte des points suivants [BOU96] :

- résistances d'accès au drain et à la source,
- modulation de la longueur effective du canal,
- variation de la mobilité en fonction des champs électriques,
- effet canal court et étroit,
- effet de substrat,
- phénomènes de conduction sous le seuil,
- modèle de conservation de la charge,
- modèle non-quasi-statique…

Deux approches de modélisation, conduisant à des modèles ayant des performances très différentes, peuvent être envisagées selon le degré de précision souhaité. La première conduisant à l'élaboration de l'ensemble des modèles de type SPICE, est fondée sur l'approximation du canal graduel. Dans ce cas, le canal du transistor est équivalent à une conductance modulée par la tension de grille et on utilise la loi d'Ohm pour calculer ces conductances non linéaires.
La deuxième approche, basée sur le calcul du potentiel de surface, est plus délicate à mettre en œuvre car elle nécessite la résolution d'équations implicites couplées. En revanche, cette deuxième approche, basée sur les équations de la physique est de loin la plus rigoureuse.

c. Modèles SPICE de première génération

L'acronyme SPICE signifie « Simulation Program with Integrated Circuit Emphasis ». Le noyau original du simulateur SPICE a été développé à l'université de Californie à Berkeley à la fin des années 60.
Le modèle SPICE de niveau 1 est à la base des modèles les plus sophistiqués qui ont été développés par la suite. L'approche mathématique utilisée et le faible nombre de paramètres de ce modèle reflètent son caractère simpliste où très peu de paramètres sont pris en compte. De plus, beaucoup d'effets comme la réduction de la mobilité ou les petites géométries ne sont pas pris en compte [FOT97]. Ce modèle de transistor est inadapté pour les technologies actuelles et ne peut être utilisé pour la conception de circuits. Il est cependant important dans un contexte historique et fournit un bon outil de compréhension pédagogique.
Le modèle SPICE de niveau 2 comble certaines lacunes du modèle de niveau 1, notamment la prise en compte des effets des petites géométries. Une autre innovation majeure concerne la prise en compte de la variation de la charge de déplétion le long du canal. L'approche utilisée a consisté à reprendre les bases du modèle antérieur et d'y ajouter de nouvelles équations et paramètres de manière à prendre en compte beaucoup plus de phénomènes physiques. Il en résulte une équation plus complexe mais plus précise du courant de drain prenant en compte les effets canaux courts, la modulation de la longueur du canal, la saturation de la vitesse des porteurs, la réduction de la mobilité... Ce qui conduit à un modèle mathématiquement très complexe.
Le modèle SPICE de niveau 3 se base sur une approche semi-empirique tout en gardant les dépendances géométriques dans ses équations, comme ses prédécesseurs. De part sa simplicité, ce modèle remporta un réel succès. En effet, ce modèle est plus simple, plus rapide et plus précis que le modèle de niveau 2 [FOT97]. De plus, les problèmes de convergence ont été corrigés. Cependant, ce modèle n'a pas été conçu pour être précis pour de grandes longueurs de canal, mais pour des longueurs minimales d'une technologie donnée. Le modèle de courant sous le seuil n'est pas physiquement réaliste et son incapacité à modéliser avec précision la conductance de sortie limite son utilisation en conception analogique.

d. Modèles de deuxième génération et de troisième génération

Une approche complètement différente a été adoptée pour cette nouvelle famille de modèles qui regroupe les modèles BSIM2 et HSPICE Level 28. Les modèles de deuxième génération insistent plus sur le conditionnement mathématique des équations du modèle de manière à réaliser des simulations de circuits plus efficaces et plus robustes.
La formulation du jeu d'équations définissant le comportement du dispositif contient les paramètres de longueur et de largeur du canal. Bien que cette formulation soit identique à la structure de base des modèles de première génération, la nouveauté de cette famille de modèles réside dans le fait que les paramètres du modèle ne sont plus considérés de manière isolés et indépendants. Considérons un paramètre quelconque X du modèle affecté d'une valeur X_0 (supposée connue). La formulation affecte à ce paramètre X, pour un transistor de longueur effective L_{eff} et de largeur effective W_{eff}, la valeur suivante [FOT97] :

$$X = X_0 + \frac{LX}{L_{eff}} + \frac{WX}{W_{eff}} \quad \text{(II. 2)}$$

Le paramètre X est dépendant de trois paramètres X_0, LX et WX. LX et WX représentent respectivement les variations en longueur et en largeur du paramètre X. La dépendance géométrique est ainsi intégrée dans tous les paramètres du modèle.

Les modèles de troisième génération sont basés sur la notion de tension de seuil, d'où leur qualificatif usuel de modèles à base de tension de seuil.
Les équations des modèles de deuxième génération sont construites sur des bases semi empiriques, voire complètement empiriques. Les paramètres de ces modèles fournissent par conséquence très peu d'informations concernant la technologie du processus de fabrication qu'ils décrivent. De plus, le nombre de paramètres utilisés est très élevé.
Les modèles BSIM3/4 et MOS Model 9 représentent l'émergence de la troisième génération de modèles. Ces modèles introduisent une description physique dans la formulation du modèle ainsi que dans le choix de ses paramètres. L'objectif premier est de lier les jeux de paramètres du modèle au processus de fabrication technologique du dispositif.

B. Le MOS Model 9

1 Structure du MOS Model 9

Le MOS modèle 9 (MM9) est un modèle très performant développé dans les années 1990 par Philips. Longtemps confidentiel, il s'est imposé comme standard de base chez STMicroelectronics, Siemens et Philips. Cependant, le MOS Modèle 9 tarde à acquérir le statut de standard mondial, face à son rival, le modèle BSIM3 de Berkeley.
Ce modèle compact est destiné à être implémenté dans des simulateurs électriques de type SPICE.
Une prédiction fine du comportement des transistors implique l'utilisation de modèles de plus en plus sophistiqués. En effet, la continuité entre les différents domaines (saturés, linéaires et bloqués) doit être assurée dans certaines régions des caractéristiques, notamment près de la tension de seuil, de manière à éviter des instabilités ou erreurs numériques lorsque le transistor opère dans des polarisations proches de ces zones de transition.
Le MOS modèle 9 fournit une description complète du transistor MOS intrinsèque. Les courants nodaux, le courant de faible avalanche, les charges affectées aux nœuds ainsi que les fluctuations dues aux bruits sont décrits.
Le MOS modèle 9 comprend 68 paramètres divisés en trois catégories principales :

- les paramètres décrivant le transistor de référence,
- les paramètres de dépendance en géométrie (longueur et largeur du canal),
- les paramètres de dépendance en température.

Le MOS modèle 9 introduit des fonctions de lissage numériques dans la structure du modèle, ces fonctions de lissage ont deux objectifs. Le premier est de permettre la continuité des équations au niveau des points de transition. Le second est de développer un seul modèle d'équation valide dans toutes les régions. Ainsi, il n'existe dans ce modèle qu'une seule expression du courant de drain qui est valable pour toutes les zones de fonctionnement du transistor, comme nous le verrons par la suite.
Cette partie est consacrée à la présentation des principales caractéristiques et équations du MOS Modèle 9 et non pas à une description laborieuse des équations du modèle, aisément accessibles à partir des références [PHI01].

2 Equations du modèle

a. Paramètres géométriques du modèle

La philosophie du MOS Modèle 9 est de rendre les principales variables du modèle dépendantes de la longueur L et de la largeur W du canal. Cette propriété est primordiale pour

notre étude puisque les variations du processus de fabrication se traduisent, dans bien des cas, par des variations géométriques sur le modèle de simulation.
Le MOS Model 9 modélise la dépendance en géométrie des différents paramètres électriques de la manière suivante : considérons un paramètre électrique X du modèle qui prend une valeur X_R (supposée connue) pour le transistor de référence dont la longueur effective et la largeur effective du canal sont $L_{eff,ref}$ et $W_{eff,ref}$. Alors, la formulation MOS Model 9 affecte à ce paramètre X, pour un transistor de longueur effective L_{eff} et de largeur effective W_{eff}, la valeur suivante :

$$X = X_R + SLX\left(\frac{1}{L_{eff}} - \frac{1}{L_{eff,ref}}\right) + SWX\left(\frac{1}{W_{eff}} - \frac{1}{W_{eff,ref}}\right) \quad \text{(II. 3)}$$

Deux nouveaux paramètres notés *SLX* et *SWX* apparaissent. *SLX* est le paramètre de dépendance en longueur de X et *SWX* le paramètre de dépendance en largeur du paramètre X.
Le MOS Model 9 offre une modélisation rigoureuse des différents effets liés à la géométrie dont les principaux sont les suivants :

- modulation de la longueur effective du canal,
- phénomènes liés au partage des charges comme les effets canaux courts ou le DIBL (Drain Induced Barrier Lowering) [GHI99].

Ces effets engendrent une variation de la tension de seuil, une dégradation de la pente sous le seuil qui se traduit par une augmentation du courant de consommation I_{off}.
Le phénomène de modulation de la longueur effective du canal se manifeste en régime de saturation. Lorsque la tension de drain est supérieure à la tension de saturation du transistor notée V_{dsat}, au niveau de la zone de drain, la zone d'inversion tend à disparaître et la zone de désertion à s'agrandir. Dans cette région la composante longitudinale du champ électrique devient non négligeable et le point de pincement du canal se déplace vers la source pour des tensions de drain grandissantes. Cela se traduit au niveau de la caractéristique électrique $I_d(V_{gc})$ par une pente positive de la caractéristique dans le régime de saturation.
Le MOS Model 9 prend en compte l'effet de modulation de la longueur effective du canal à l'aide de deux paramètres nommés ALP (coefficient de modulation de la longueur du canal) et VPR (prise en compte de la tension d'Early).
L'effet canal court est dû à une variation du potentiel de surface le long du canal, notamment au niveau des jonctions de drain et de source. L'empiètement des zones de charges d'espace de la source et du drain sur la zone de désertion (générée et contrôlée par le potentiel de grille) entraîne une perte du monopole de contrôle de la grille et une diminution de la tension de seuil.
L'effet DIBL s'ajoute à l'effet canal court et accentue la diminution de la tension de seuil pour de faibles longueurs de canal. Cet effet engendre des non linéarités pour de fortes polarisations côté drain. L'augmentation de la tension de drain diminue la hauteur de la barrière de potentiel et augmente la zone de désertion côté drain. Dans la formulation MOS Model 9, la variation de la tension de seuil en fonction de la tension drain est mesurée avec le coefficient DIBL noté GAMOOR. La variation de la pente du courant de drain sous le seuil est prise en compte par le paramètre MOR. Par ailleurs, pour garantir une transition continue entre l'inversion faible et l'inversion forte, on introduit le paramètre ZET1R dans l'expression du courant de drain.

b. Equation du courant de drain

La seconde philosophie du MOS Model 9 est de développer un seul modèle d'équation valide dans toutes les régions de fonctionnement du transistor. Ainsi, il n'existe dans ce modèle qu'une seule expression du courant de drain qui est valable pour toutes les zones de fonctionnement du transistor :

$$Ids = \beta.G3.\left(\delta \frac{V_{GT3}.V_{DS1} - \frac{1+\delta_1}{2}(V_{DS1})^2}{(1+\theta_3.V_{DS1}).(1+\theta_1(V_{GS1}-V_{TH})+\theta_2(\sqrt{\Phi_B - V_{BS}} - \sqrt{\Phi_B}))} \right) \quad \text{(II. 4)}$$

Dans cette dernière équation, le paramètre θ_1 par exemple, que nous nommerons THE1 dans l'équation II.5 est calculé comme suit :

$$THE1 = \theta_{1R} + S_{L,\theta_1}\left(\frac{1}{L_{eff}} - \frac{1}{L_{eff,ref}}\right) + SWTH1\left(\frac{1}{W_{eff}} - \frac{1}{W_{eff,ref}}\right) \quad \text{(II. 5)}$$

Dans cette dernière équation, THE1 est un paramètre classique du modèle et SWTH1 un paramètre qui traduit les dépendances géométriques (largeur de canal) du paramètre THE1. θ_{1R} correspond à la valeur de référence du paramètre THE1.

c. Modèle de mobilité

Dans la zone de fonctionnement linéaire, la mobilité μ_n, considérée dans l'expression du courant de drain, est la mobilité idéale d'un électron supposé éloigné de l'interface oxyde-semiconducteur. En réalité, le champ vertical E_T créé par la polarisation de la grille intervient dans l'expression de la mobilité μ_n de la manière suivante :

$$\mu_n = \frac{\mu_0}{1+\alpha.E_T} \quad \text{(II. 6)}$$

A partir de cette expression, nous déterminons une mobilité effective qui est fonction de la polarisation de grille en introduisant le paramètre θ_1 (que l'on retrouve dans l'équation II.4) qui modélise la réduction de la mobilité due au champ vertical :

$$\mu_{eff} = \frac{\mu_0}{1+\theta_1.(V_{GS}-V_{TH})} \quad \text{(II. 7)}$$

La mobilité est aussi affectée par la polarisation de substrat via un paramètre θ_2 (que l'on retrouve aussi dans l'équation II.4), qui prend en compte la réduction de la mobilité en fonction de la polarisation de substrat. L'expression de la mobilité devient alors :

$$\mu_{eff} = \frac{\mu_0}{1+\theta_1.(V_{GS}-V_{TH})+\theta_2.(\sqrt{(\Phi_B - V_{BS})} - \sqrt{\Phi_B})} \quad \text{(II. 8)}$$

Cette dernière expression de la mobilité ne fait intervenir que le champ vertical et la polarisation du substrat. Cette expression n'est donc valable que pour de faibles tensions de drain. Lorsque la tension de drain augmente, la mobilité est affectée par la composante latérale du champ électrique qui règne au niveau du canal (la vitesse de saturation des porteurs V_{max} est voisine de 100.000 m/s lorsque le champ latéral atteint une valeur critique). Plusieurs expressions empiriques ont été établies pour prendre en compte ce phénomène. L'expression la plus répandue étant la suivante :

$$\mu_{eff\,1} = \frac{\mu_{eff}}{1 + \mu_{eff} \cdot \frac{V_{DS}}{v_{max}.L}} = \frac{\mu_{eff}}{1 + \theta_3 . V_{DS}} \quad \text{(II. 9)}$$

La mobilité μ_{eff1} est la mobilité résultant de la prise en compte des effets du champ latéral, du champ vertical et de la polarisation de substrat. μ_{eff} étant l'expression de la mobilité ne tenant compte que des effets du champ vertical et de la polarisation de substrat.
La réduction de la mobilité due au champ latéral est modélisée par le paramètre θ_3. Ce paramètre est relié à la vitesse d'entraînement des porteurs par la formule suivante :

$$\theta_3 = \frac{\mu_{eff}}{v_{max}.L} \quad \text{(II. 10)}$$

d. Modèle de la charge

Comme de nombreux modèles de charge, le développement du MOS Model 9 est basé sur l'équation de conservation de la charge suivante :

$$Q_g + Q_{depl} + Q_{inv} = 0 \quad \text{(II. 11)}$$

La charge totale de déplétion Q_{depl} et la charge d'inversion Q_{inv} sont calculées à partir des équations du modèle. La charge de grille Q_g est alors obtenue via l'expression :

$$Q_g = -(Q_{depl} - Q_{inv}) \quad \text{(II. 12)}$$

La charge en régime d'inversion Q_{inv} (variable en fonction du régime de fonctionnement) est composée de la charge associée à l'électrode de drain et de la charge associée à l'électrode de source. Le mode de répartition de la charge, basé sur des principes physiques, obéit à une méthode couramment utilisée dans les autres types de modèles [SAK90].
Des fonctions de lissage sont ensuite incorporées dans les équations de manière à rendre le modèle continu dans toutes les zones de fonctionnement.

C. Modélisation du transistor MOS à grille flottante

1 Modélisation statique : le modèle capacitif

Les mémoires non volatiles de type FLOTOX possèdent des propriétés spécifiques qui doivent être prises en compte pour leur compréhension et leur modélisation. Du point de vue des caractéristiques électriques ($I_d(V_{cg})$, $I_d(V_{ds})$, …), il n'y a pas de différences fondamentales entre un transistor MOS conventionnel et une mémoire non volatile. Cependant, la modélisation de l'effet mémoire nécessite la prise en compte de certaines considérations évoquées au Chapitre I (cf. §I.A.4). En effet, la présence d'une grille flottante entourée d'un diélectrique entraîne des conséquences importantes du point de vue de la modélisation. Ces conséquences seront développées dans cette partie.
De part la structure des mémoires non volatiles, la grille flottante n'est pas accessible de manière directe. La tension de la grille flottante est contrôlée par un effet de couplage capacitif des tensions appliquées sur les nœuds externes de la mémoire. Les mémoires non volatiles sont souvent modélisées par un circuit équivalent appelé « modèle capacitif ». Ce modèle, présenté figure II.5, fait apparaître les principales capacités présentent dans une structure mémoire double polysilicium à grille flottante : C_g représente la capacité totale entre la grille flottante et la grille de contrôle, C_s et C_d sont les capacités de source et de drain et C_k représente la capacité entre la grille flottante et le canal du transistor mémoire. Notons que C_d est la capacité relative à la région définie par l'oxyde tunnel.

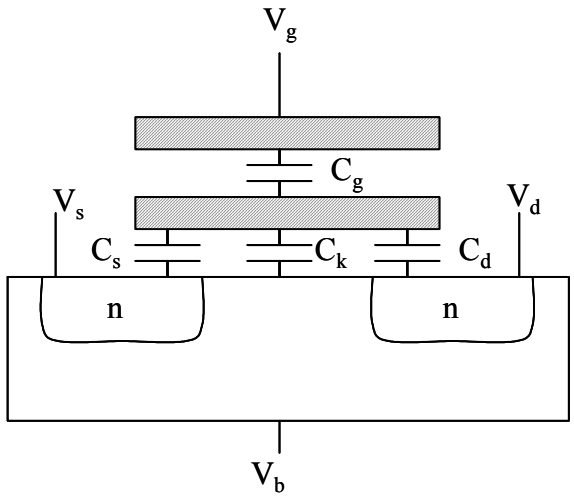


Figure II. 5 Modèle capacitif.

La capacité totale relative à la grille flottante est :

$$C_t = C_k + C_s + C_d + C_g \quad \text{(II. 13)}$$

Deux coefficients de couplage importants apparaissent :

- Le coefficient de couplage de la grille de contrôle :

$$\alpha_g = \frac{C_g}{C_t} \quad \text{(II. 14)}$$

- Le coefficient de couplage de la zone de drain :

$$\alpha_d = \frac{C_d}{C_t} \quad \text{(II. 15)}$$

Ces coefficients de couplage ont une importance primordiale lors des phases de conception et de modélisation puisqu'ils ont un impact direct sur la qualité de la programmation de la mémoire.
Le modèle capacitif peut être utilisé pour le calcul du potentiel de grille flottante à partir de la connaissance des tensions appliquées sur les nœuds externes et de la charge initiale de la grille flottante notée Q_{gf}. Le potentiel de grille flottante ainsi calculé peut être ensuite substitué au potentiel de grille d'un transistor MOS conventionnel. Ainsi, à partir des équations du transistor MOS, les caractéristiques électriques du transistor mémoire peuvent être extraites. Réciproquement, ce modèle capacitif peut être utilisé pour le calcul de certains paramètres du modèle comme les coefficients de couplage α_g et α_d à partir de mesures sur le transistor MOS équivalent (transistor mémoire dont la grille flottante est accessible sous forme de nœud externe).
Le calcul du potentiel de grille flottante V_{GF} du transistor mémoire peut être obtenu en utilisant l'équation de conservation de la charge au niveau de la grille flottante, associée au modèle capacitif.

a. Calcul du potentiel de grille flottante pour $Q_{gf} = 0$

Si l'on considère le cas où aucune charge n'est piégée dans la grille flottante, c'est-à-dire $Q_{gf}=0$, l'équation de conservation de la charge est donnée par l'équation II.16 :

$$Q_{gf} = 0 = C_g(V_{GF} - V_g) + C_s(V_{GF} - V_s) + C_d(V_{GF} - V_d) + C_k(V_{GF} - V_b) \quad \text{(II. 16)}$$

V_g étant le potentiel appliqué sur la grille de contrôle et V_s, V_d et V_b les potentiels de source, de drain et de substrat. En faisant intervenir les coefficients de couplage dans cette dernière équation (équations II.14 et II.15), on obtient :

$$V_{GF} = \alpha_g.V_g + \alpha_d.V_d + \alpha_s.V_s + \alpha_b.V_b \quad \text{(II. 17)}$$

α_J étant le coefficient de couplage relatif à l'électrode J, où J peut représenter l'électrode de grille, de drain ou de source. Si l'électrode de source et le substrat sont reliés à la masse, l'équation II.17 se simplifie :

$$V_{GF} = \alpha_g.(V_{gs} + \frac{\alpha_d}{\alpha_g}.V_{ds}) = \alpha_g.(V_{gs} + f.V_{ds}) \quad \text{(II. 18)} \qquad \text{Avec,} \quad f = \frac{\alpha_d}{\alpha_g} = \frac{C_d}{C_g} \quad \text{(II. 19)}$$

Les équations du transistor MOS à grille flottante peuvent être obtenues à partir des équations conventionnelles de transistor MOS en remplaçant la tension de grille V_g du transistor MOS

par le potentiel de grille flottante. De ce fait, les paramètres de la cellule mémoire, tels que la tension de seuil V_T et le facteur de conductivité β peuvent être recalculés et exprimés en fonction des coefficients de couplage :

$$V_T^{GF} = V_{T\,(\text{grille flottante})} = \alpha_g . V_{T\,(\text{grille de contrôle})} = \alpha_g . V_T^{GC} \quad \text{(II. 20)}$$

$$\beta^{GF} = \beta_{(\text{grille flottante})} = \frac{1}{\alpha_g} \beta_{(\text{grille de contrôle})} = \frac{\beta^{GC}}{\alpha_g} \quad \text{(II. 21)}$$

On retrouve ainsi les équations du transistor MOS à grille flottante en régime quadratique :

$$I_{DS} = \beta^{GC} . \left[(V_{GS} - V_T^{GC}) V_{DS} - (f - \frac{1}{2.\alpha_g}) . V_{DS}^2 \right] \quad \text{(II. 22)} \qquad |V_{DS}| < \alpha_g . \left| V_{GS} + f . V_{DS} - V_T^{GC} \right|$$

Et en régime saturé :

$$I_{DS} = \frac{\beta^{GC}}{2} . \alpha_g . (V_{GS} + f . V_{DS} - V_T^{GC})^2 \quad \text{(II. 23)} \qquad |V_{DS}| \geq \alpha_g . \left| V_{GS} + f . V_{DS} - V_T^{GC} \right|$$

Ces équations nous permettent d'observer plusieurs effets. La majorité d'entre eux sont dus au couplage capacitif entre le drain et la grille flottante, qui a pour conséquence de modifier les caractéristiques courant-tension des transistors MOS à grille flottante par rapport aux transistors MOS conventionnels :

- le transistor à grille flottante peut entrer en régime de déplétion et conduire le courant lorsque $V_{GS}<V_T$. Ceci parce que la conduction du canal peut être induite par la tension de drain. Le couplage capacitif entre le drain et la grille flottante intervient dans l'équation II.23 avec le paramètre $f.V_{DS}$,
- la région de saturation pour le transistor MOS conventionnel correspond à une région où le courant de drain est pratiquement indépendant du potentiel de drain V_{DS}. Dans le cas du transistor à grille flottante et pour les mêmes conditions de polarisation, le courant de drain continue à augmenter avec la tension de drain, en conséquence le régime de saturation ne peut être atteint,
- la limite entre le régime résistif et le régime de saturation du transistor à grille flottante est donnée par la relation :

$$|V_{DS}| = \alpha_g . \left| V_{GS} + f . V_{DS} - V_T^{GC} \right| \quad \text{(II. 24)}$$

- la transconductance dans la région de saturation est donnée par l'expression :

$$g_m = \frac{\delta I_{DS}}{\delta V_{GS}} = \alpha_g . \beta . (V_{GS} + f . V_{DS} - V_T^{GC}) \quad \text{(II. 25)}$$

la transconductance g_m est dépendante de la tension de drain V_{DS}, contrairement au transistor conventionnel où la transconductance est considéré comme relativement indépendante de la tension de drain dans la région de saturation;

- le coefficient de couplage capacitif f dépend des capacités C_d et C_g uniquement, et sa valeur peut être vérifiée, dans le régime de saturation par l'expression :

$$f = -\frac{\delta V_{GS}}{\delta V_{DS}} \quad \text{(II. 26)}$$

De nombreuses techniques existent pour extraire les coefficients de couplage capacitif à partir de mesures de caractéristiques électriques simples [PRA87] [WAD80]. Les méthodes les plus largement répandues sont la technique linéaire de la tension de seuil, la méthode de la pente sous le seuil ou encore la technique de la transconductance [WON92] [CHO94].
Ces méthodes exigent la mesure de paramètres électriques sur une cellule mémoire classique et sur une cellule mémoire atypique appelée « dummy cell » où la grille flottante et la grille de contrôle du transistor mémoire sont connectées. Par comparaison des mesures effectuées sur ces deux structures, les coefficients de couplage peuvent être déterminés.

b. Calcul du potentiel de grille flottante pour $Q_{gf} \neq 0$

Dans le cas où la charge stockée dans la grille flottante Q_{gf} est non nulle, l'équation de conservation de la charge et le potentiel de grille flottante V_{GF} apparaissent sous la forme suivante :

$$Q_{gf} = C_g(V_{GF} - V_g) + C_s(V_{GF} - V_s) + C_d(V_{GF} - V_d) + C_k(V_{GF} - V_b) \quad \text{(II. 27)}$$

$$V_{GF} = \alpha_g.V_g + \alpha_d.V_d + \alpha_s.V_S + \alpha_b.V + \frac{Q_{gf}}{C_t} = \alpha_g.V_g + \alpha_d.V_d + \frac{Q_{gf}}{C_t} \quad \text{(II. 28)}$$

A partir des équations II.28 et II.20, on obtient :

$$V_T{}^{CG} = V_{T(grille\ contrôle)} = \frac{1}{\alpha_g}.V_T{}^{GF} - \frac{Q_{gf}}{C_t.\alpha_g} = \frac{1}{\alpha_g}.V_T{}^{GF} - \frac{Q_{gf}}{C_g} \quad \text{(II. 29)}$$

Cette dernière équation montre la dépendance de la tension de seuil au niveau de la grille de contrôle $V_T{}^{CG}$ du transistor mémoire par rapport à la charge Q_{gf} de la grille flottante. Le décalage de la tension de seuil ΔV_T est proportionnel à la charge emmagasinée dans la grille flottante, il est donné par la différence entre l'équation II.29 et l'équation II.20 :

$$\Delta V_T = V_T - V_{T(Q_{gf}=0)} = V_T - V_{T0} = V_T - \alpha_g . V_T = -\frac{Q_{fg}}{C_g} \quad \text{(II. 30)}$$

V_{T0} représente la tension de seuil pour une charge de grille flottante nulle. Cette équation montre clairement que la quantité de charges injectée dans la grille flottante engendre un décalage de la tension de seuil du transistor mémoire.

2 Mécanisme d'injection de charges par effet tunnel

Le mécanisme d'effet tunnel Fowler-Nordheim (FN) a lieu quand un champ électrique élevé est appliqué aux bornes d'un oxyde mince. Dans ces conditions, les bandes d'énergie de la région de l'oxyde sont très raides, assurant une forte probabilité de passage de la barrière d'énergie pour les électrons.
Une description détaillée du phénomène d'injection de charges par effet tunnel a été présentée au Chapitre I (cf. §I.B.1). Rappelons que le mécanisme d'injection par effet tunnel est largement répandu dans les mémoires non volatiles, en particulier dans les EEPROM. Ce choix étant justifié par les raisons suivantes :

- l'effet tunnel est un mécanisme électrique pur,
- le niveau des courants induits est très faible et permet ainsi la génération des tensions d'alimentation requises pour tous les fonctionnements,
- il permet d'obtenir des temps de programmation inférieurs à 1 ms et des temps de rétention supérieurs à 10 ans, ce qui est fondamental pour toutes les technologies non volatiles.

De plus, il a été vu (cf. Chapitre I, §I.B.1a) que la densité de courant tunnel pouvait s'écrire sous la forme simple :

$$J = \alpha . E^2 . e^{\left(\frac{-\beta}{E}\right)} \quad \text{(II. 31)}$$

où α et β sont des coefficients qui dépendent du champ électrique aux bornes de l'oxyde.

3 Modélisation dynamique

Le transfert de charge dans la grille flottante conduit à un changement de son potentiel. Ce changement de potentiel induit une variation du champ électrique aux bornes des oxydes entourant la grille flottante. Ce qui provoque une diminution ou une augmentation du courant d'injection tunnel. Le courant d'injection n'est pas constant durant les phases de programmation de la mémoire, il varie rapidement en fonction du temps et du niveau des potentiels appliqués aux nœuds externes de la mémoire.
L'élaboration d'un modèle dynamique capable de décrire les caractéristiques transitoires de la mémoire se base sur l'équation de la charge de la grille flottante II.32 :

$$\frac{dQ_{gf}(t)}{dt} = \int_{A_{gf}} J_{gf}(\mathrm{t}).\mathrm{dA} \quad \text{(II. 32)}$$

Dans cette expression, l'intégrale calculée sur l'aire totale A_{gf} de la grille flottante peut être remplacée par une sommation des flux de densité de courant J_i (qui dépendent du champ électrique) traversant les différents oxydes 'i' qui entourent la grille flottante :

$$\frac{dQ_{gf}(t)}{dt} = \int_{A_i} J_i\left[E_i(t)\right] \quad \text{(II. 33)}$$

En intégrant cette dernière équation dans le temps, l'expression de la charge accumulée dans la grille flottante peut être obtenue par l'équation II.34 :

$$Q_{gf}(t) = \int_0^t \sum_{A_i} J_i\left[E_i(t)\right]\mathrm{dt} \quad \text{(II. 34)}$$

Les différents courants traversant les diélectriques sont fonction du temps, puisque le champ électrique aux bornes des oxydes évolue avec le temps.
De ce fait, l'équation II.34 ne peut être directement résolue puisque le champ électrique E_i à travers les oxydes dépend du potentiel de grille flottante qui dépend à son tour de la charge de grille flottante Q_{gf}. La relation entre le potentiel de grille flottante et la charge de grille flottante Q_{gf} étant donnée par l'équation II.28.
La connaissance des différents potentiels appliqués aux noeuds externes du transistor mémoire permet de déterminer les différents champs électriques E_i aux bornes de chaque diélectrique.
Dans le cas des mémoires non volatiles de type EEPROM, la connaissance de ces champs électriques, associés à un modèle d'injection de charges, met en évidence un transfert de charge de type FN à travers l'oxyde tunnel. En effet, pour les cellules EEPROM de type FLOTOX, une couche d'oxyde mince sépare le drain de la grille flottante. Sous l'effet d'un champ électrique intense E_{tun} de l'ordre de 10MV/cm, les électrons passent par effet tunnel à travers cet oxyde mince de surface S_{tun}. Ce mode de transfert de charges étant prédominant, l'équation II.35 devient :

$$Q_{gf}(t) = \int_0^t \sum_{S_{tun}} J_{tun}\left[E_{tun}(t)\right].dt \quad \text{(II. 35)}$$

En prenant en compte la charge de grille flottante initiale Q_{fg0} et la variation du champ électrique à travers l'oxyde tunnel, le potentiel de grille flottante peut être calculé à chaque instant par la résolution des équations suivantes :

$$Q_{gf}(t) = Q_{gf\,0} - \int_0^t I_{tun}\left[E_{tun}(t)\right].dt \quad \text{(II. 36)}$$

Avec : $$I_{tun} = S_{tun}.J_{tun}\left[E_{tun}(t)\right] = S_{tun}.\alpha.E_{tun}(t)^2.e^{\left(\frac{-\beta}{E_{tun}(t)}\right)} \quad \text{(II. 37)}$$

II. Modélisation du point mémoire EEPROM

A. La cellule EEPROM « F6DP 7.7µm^2 »

L'approche utilisée pour développer le modèle EEPROM se base sur l'équation de neutralité de la charge au niveau de la grille flottante. La résolution de cette équation détermine le potentiel de grille flottante à partir duquel toutes les variables du transistor mémoire peuvent être calculées en utilisant la formulation MOS Model 9.

1 Caractéristiques de la cellule

La figure II.6 représente une coupe du point mémoire EEPROM. Le point mémoire EEPROM est constitué de deux transistors : un transistor de sélection, sur la droite de la figure II.6, qui va transmettre la haute tension durant la phase d'écriture et le transistor mémoire canal n, constitué de deux couches de polysilicium. La première couche, au dessus du canal et entourée de diélectrique représente la grille flottante. La deuxième couche en polysilicium représente la grille de contrôle du transistor mémoire. Dans le modèle de mémoire EEPROM développé, le transistor formé par l'électrode de drain, l'électrode de source et la grille flottante du transistor mémoire est modélisé à partir d'un transistor MOS Model 9.

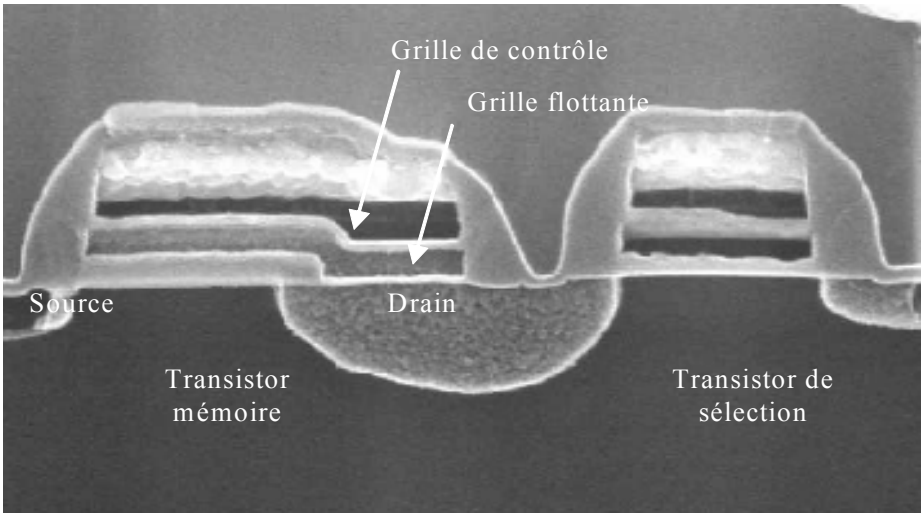
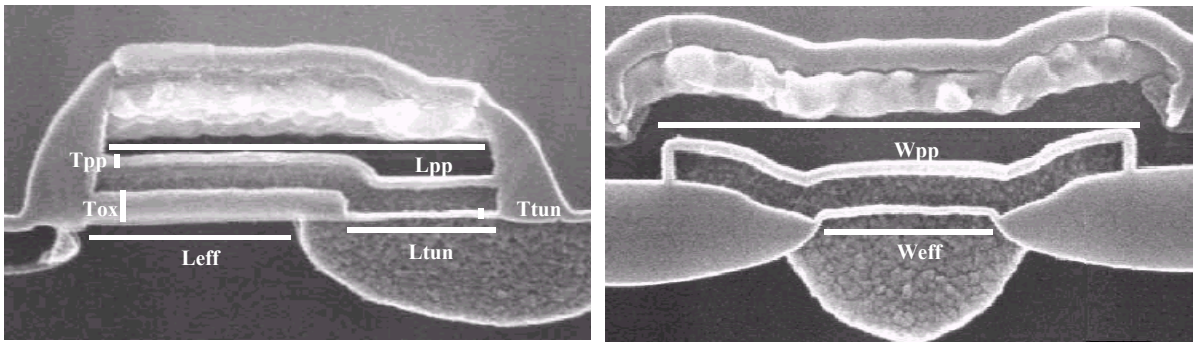


Figure II. 6 Vue en coupe du point mémoire EEPROM.

Les paramètres importants que l'on peut associer au transistor mémoire sont décrits figure II.7. T_{pp}, T_{tun} et T_{ox} représentent les épaisseurs respectives de l'oxyde ONO, de l'oxyde tunnel et de l'oxyde de champ au dessus du canal. L_{eff} représente la longueur effective du canal et L_{tun} la longueur de la fenêtre tunnel. Sur la coupe transversale du transistor mémoire présentée figure II.7b, on distingue la largeur W_{eff} du transistor mémoire ainsi que la largeur de l'oxyde ONO W_{pp}.
La connaissance de ces paramètres géométriques peut nous permettre de calculer la capacité inter polysilicium C_{pp}, la capacité tunnel C_{tun} et des coefficients de couplage associés.
Ces paramètres seront utilisés lors du diagnostic de défauts géométrique de la cellule EEPROM.



(a) Coupe longitudinale (b) Coupe transversale

Figure II. 7 Paramètres géométriques du transistor mémoire

2 Modélisation statique de la cellule EEPROM

La loi de Gauss appliquée à la surface fermée entourant la grille flottante permet le calcul de la charge de grille flottante Q_{gf} par la résolution de l'équation suivante :

$$Q_{gf} = C_{pp}(V_{GF} - V_g) + C_{tun}(V_{GF} - V_d) + Q_g + Q_{gf0} \quad \text{(II. 38)}$$

Q_{gf0} représente la charge initiale de grille flottante et Q_g la charge de grille du transistor MOS Model 9 vérifiant l'équation de conservation de la charge II.11.
Q_g est calculée en utilisant le modèle de charge contenu dans la formulation MOS Model 9. Cette charge dépend du potentiel de grille flottante V_{GF} (mais aussi du régime de fonctionnement du transistor), par conséquent, l'équation II.38 est une équation implicite de la variable V_{GF}.
La connaissance du potentiel de grille flottante V_{GF}, appliqué sur la grille du transistor MOS Model 9 permettra de calculer toutes les autres variables du modèle.

3 Modélisation dynamique de la cellule EEPROM

L'approximation non quasi statique a été utilisée pour le développement du modèle dynamique. En régime dynamique, le potentiel de grille flottante est calculé par la résolution de l'équation implicite suivante :

$$Q_{gf}(t) - Q_{gf\,0} - \int_0^t I_{tun}\left[E_{tun}(t)\right]dt = 0 \quad \text{(II. 39)}$$

L'expression du courant d'injection FN étant utilisée sous sa forme simple :

$$I_{tun} = \alpha . S_{tun} E_{tun}{}^2 . e^{\left(\frac{-\beta}{E_{tun}}\right)} \quad \text{(II. 40)}$$

S_{tun} représente la surface de l'oxyde tunnel ($S_{tun} = L_{tun} .W_{eff}$) et E_{tun} le champ électrique aux bornes de l'oxyde tunnel.
La figure II.8 montre la partie du transistor mémoire EEPROM modélisée par le MOS Model 9. Ce modèle de cellule mémoire prend donc en compte, par définition, l'influence de tous ses paramètres géométriques sur ses caractéristiques électriques et donc sur sa tension de seuil. Le comportement dynamique du modèle est représenté par le courant d'injection I_{FN}.
Ce modèle peut être utilisé pour des simulations statiques aussi bien que transitoires et fournit des temps de simulations relativement courts. Cette dernière caractéristique permettra d'utiliser ce modèle pour la conception de matrices de cellules mémoires.

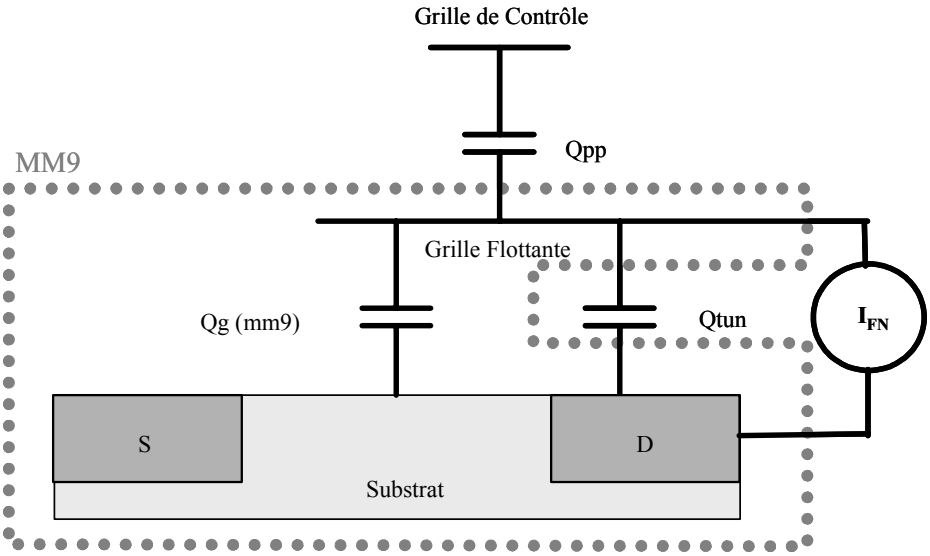


Figure II. 8 Schéma équivalent d'une cellule EEPROM.

B. La cellule EEPROM : modèle électrique

1 Les simulateurs de circuits

Les simulateurs électriques analogiques tel que ELDO permettent de simuler différents circuits composés d'éléments tels que des transistors, des diodes, des résistances, des capacités et des inductances. Les modèles de ces composants sont le plus fréquemment inclus dans le code du simulateur et la création d'un circuit se fait en composant une « netlist » (qui décrit le circuit formé par tous ses éléments de base).
Pour simuler un circuit, le simulateur forme un système d'équation en utilisant la loi de Kirchoff à chaque nœud du circuit. A partir de ce type d'équations, il existe plusieurs modes de calcul qui dépendent du type d'analyse choisi. On distingue :

- l'analyse statique DC ou point de fonctionnement, qui correspond à la résolution d'un système d'équations différentielles. Dans la majorité des cas, la présence d'éléments non linéaires impose l'utilisation d'algorithmes qui vont linéariser les équations autour d'un point d'équilibre de manière itérative;
- l'analyse transitoire permet d'analyser le circuit au cours du temps en découpant l'intervalle de temps en un nombre fini de points et en résolvant le système d'équations non linéaires à chaque pas de temps;
- l'analyse harmonique AC est précédée d'une analyse DC destinée à calculer le point de fonctionnement de tous les composants du circuit. Ensuite, la

linéarisation se fait en calculant les fonctions de transfert complexes des éléments non linéaires, en les représentant par un modèle linéaire valable pour des « petits signaux », à proximité immédiate du point de fonctionnement. Enfin, le simulateur calcule les tensions et courants complexes pour chaque valeur de fréquence demandée pour l'analyse AC;

- il est aussi possible d'effectuer des analyses statistiques suivant des dispersions de paramètres de composants avec l'analyse Monte-Carlo.

La majorité des simulateurs électriques sont basés sur la loi de Kirchoff qui est elle-même basée sur la loi de conservation de l'énergie qui peut également s'appliquer à d'autres domaines tels que la thermique ou la mécanique.

2 Modélisation comportementale HDLA de la cellule EEPROM

Le modèle de cellule mémoire EEPROM a été développé en langage HDLA. La structure générale d'un programme HDL-A est la suivante :

```
ENTITY nom IS
        Description de l'interface (PORT, PIN,…) ;
END ENTITY nom ;

ARCHITECTURE a OF nom IS
        Déclaration (CONSTANT, SIGNAL, VARIABLE, STATE,…) ;
BEGIN
        RELATION
                PROCEDURAL FOR… =>
                Calcul explicite sur les tensions, courants et variables d'état.
                EQUATION variable_d'état FOR… =>
                Calcul implicite sur les variables d'état.
        END RELATION ;
END ARCHITECTURE a ;
```

Ce type de langage de description de composants analogiques permet de modéliser un composant ou un circuit intégré en vue de leur validation par simulation. La création de modèles de composants se fait par la définition des relations mathématiques entre des variables (courants, potentiels, flux…) et leurs entrées et sorties. Des langages de description comme HDL-A, MAST et VERILOG-A sont destinés à créer des modèles fonctionnels de haut niveau d'abstraction pour la simulation.
Il y a toujours deux façons d'aborder un langage : « en bloc » ou « par analogie ». Le langage HDL-A utilise une approche en bloc comme le VHDL.
Une description en langage VHDL comporte une ou plusieurs bibliothèques dans lesquelles sont stockées des unités de conception. Les unités de conception peuvent être de cinq catégories différentes :

- le bloc entité (« ENTITY ») qui décrit les relations entre le modèle et l'extérieur, c'est-à-dire le nombre et la nature des entrées et sorties;
- le bloc architecture (« ARCHITECTURE ») qui décrit le comportement du modèle présenté dans le bloc entité. Les relations du bloc architecture permettent de lire et d'écrire les grandeurs définies dans le bloc entité et de faire des opérations mathématiques, soit directement entre ces grandeurs, soit par l'intermédiaire de variables locales;

- le bloc configuration (« CONFIGURATION ») est l'unité qui décrit la correspondance entre les composants déclarés dans les architectures et des architectures précises d'entités;
- le bloc spécification de paquetage (« PACKAGE ») comprend la description des objets, des types et des sous-programmes accessibles du paquetage;
- le corps de paquetage (« PACKAGE BODY ») contient le code des sous-programmes proposés par le paquetage et éventuellement de sous-programmes d'usage interne au paquetage.

En plus des blocs définis dans la syntaxe « VHDL » la partie architecture de la description HDL-A comporte un bloc « RELATION » qui décrit le comportement des variables associées aux nœuds du modèle. Seul un des champs (tension ou courant) lié au nœud peut être évalué, l'autre étant déterminé par la relation de Kirchoff. La première partie de ce bloc (« PROCEDURAL FOR… ») décrit les relations explicites qui lient les tensions et les courants des nœuds.
La deuxième partie (« EQUATION ») est réservée aux relations implicites entre ces grandeurs.
Les modèles décrits en langage HDL-A sont utilisables par le simulateur de circuit ELDO pour être associés à d'autres composants. Ainsi le modèle de transistor à grille flottante développé pourra être utilisé de manière à créer une matrice de cellules mémoires à laquelle on pourra associer tous les éléments périphériques comme les circuits de décodage des adresses par exemple.

3 Validation du modèle

La validation du modèle de cellule mémoire EEPROM décrit en langage HDL-A a été réalisée en utilisant des structures de test embarquées dans les lignes de découpe de la tranche de silicium. La structure de test présentée figure II.9 est constituée de deux transistors : un transistor de sélection haute tension et un transistor à grille flottante aussi appelé « sense transistor ». Ces deux transistors sont simulés à partir de cartes modèles MOS Model 9 fournies par la société ST-Microelectronics.
Les paramètres spécifiques au transistor mémoire comme l'épaisseur de l'oxyde tunnel ou l'épaisseur de l'oxyde inter poly silicium sont des paramètres du modèle qui dépendent de la technologie considérée.
Concernant l'extraction des paramètres Fowler-Nordheim α et β, des mesures basées sur des techniques I (V) et C (V) ont été utilisées [HAR00].

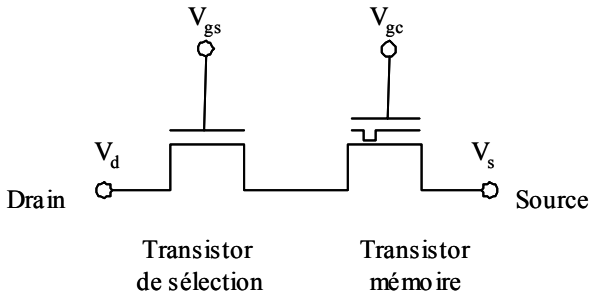


Figure II. 9 Structure de test.

La figure II.10 est une représentation des tensions de seuil obtenues pour différentes tensions de drain V_d et de grille V_{gc} (le transistor de sélection étant rendu passant). La comparaison entre les résultats obtenus par simulation et les mesures expérimentales démontre l'exactitude de modèle EEPROM (pour cet exemple de validation).

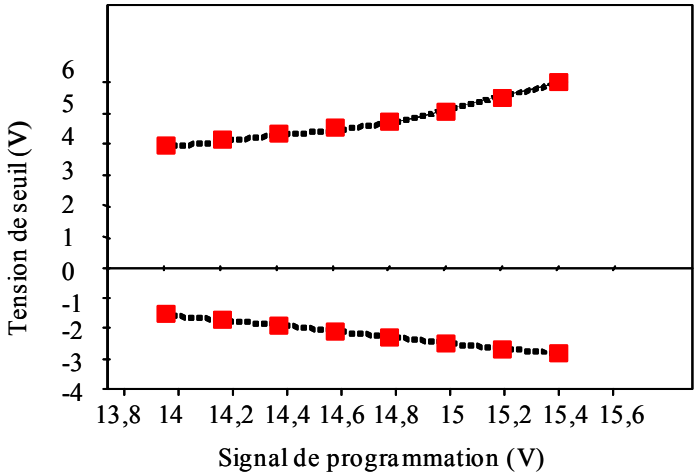


Figure II. 10 Variation de la fenêtre de programmation : simulation (---), mesure (■).

4 Résultats de simulation

Les résultats de simulation présentés dans cette partie sont obtenus avec la même carte modèle. Les simulations transitoires sont obtenues en utilisant les valeurs nominales des paramètres géométriques du transistor mémoire. Dans le cas des simulations statiques, des mesures effectuées en utilisant des géométries de cellules mémoires différentes mettent en évidence l'impact de la variation de certains paramètres géométriques sur les caractéristiques $I_d(V_{gc})$ du transistor mémoire [POR02].

a. Simulations transitoires

Les signaux de programmation V_{gc} et V_d appliqués sur la structure de test de la figure II.9 sont présentés figure II.11a durant deux cycles de programmation (effacement et écriture). Cette figure présente aussi la variation du potentiel de grille flottante V_{gf}.
Durant la phase d'effacement, correspondant à l'application de la haute tension (~ 14 V) sur la grille du transistor mémoire, le potentiel de grille flottante évolue de la même manière que le potentiel V_{gc} jusqu'à l'apparition d'un courant d'injection tunnel présenté figure II.11b.
La phase d'effacement correspondant à l'injection d'un courant tunnel Fowler-Nordheim négatif et entraîne une diminution progressive du potentiel de grille flottante. Après la phase d'effacement (V_{gc} = 0), le potentiel de grille flottante V_{gf} est négatif, ce qui correspond au piégeage de charges dans la grille flottante du transistor mémoire.
Un comportement similaire est constaté durant la phase d'écriture, avec cette fois un courant tunnel positif qui va décharger la grille flottante et engendrer une augmentation du potentiel de grille flottante V_{gf} à la fin de la phase d'écriture (V_d = 0).
Il est à noter que le potentiel de drain V_d appliqué à la structure de test est transmis sur le drain du transistor mémoire en commandant le potentiel de grille V_{gs} du transistor de sélection (Figure II.10).

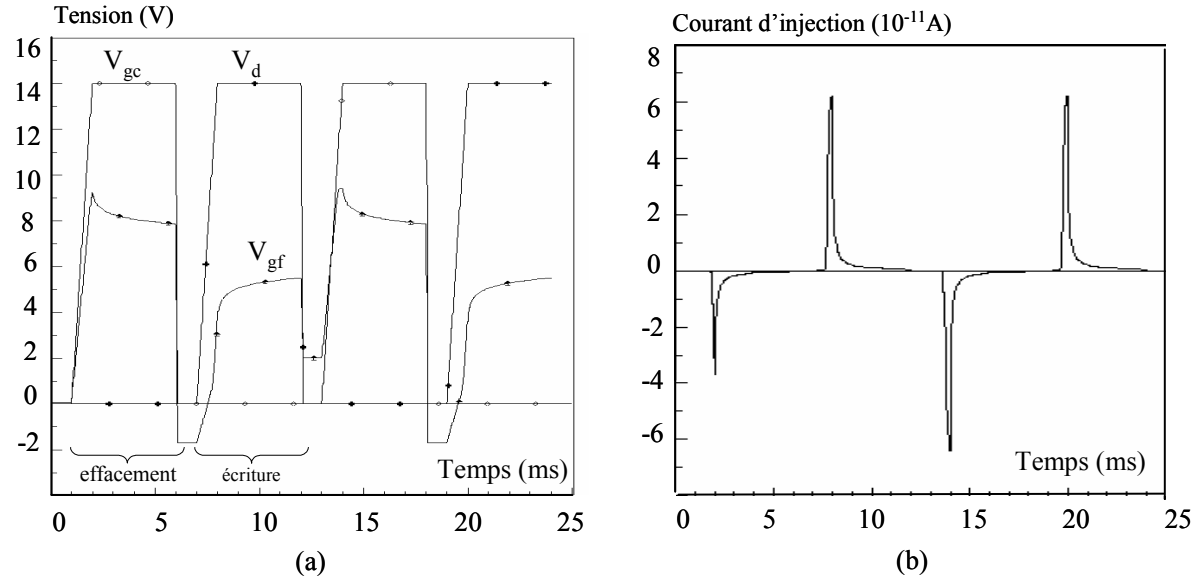


Figure II. 11 Variation des signaux de commande V_d et V_{gc}, du potentiel de grille flottante V_{gf} (a) et du courant tunnel (b) durant deux cycles de programmation.

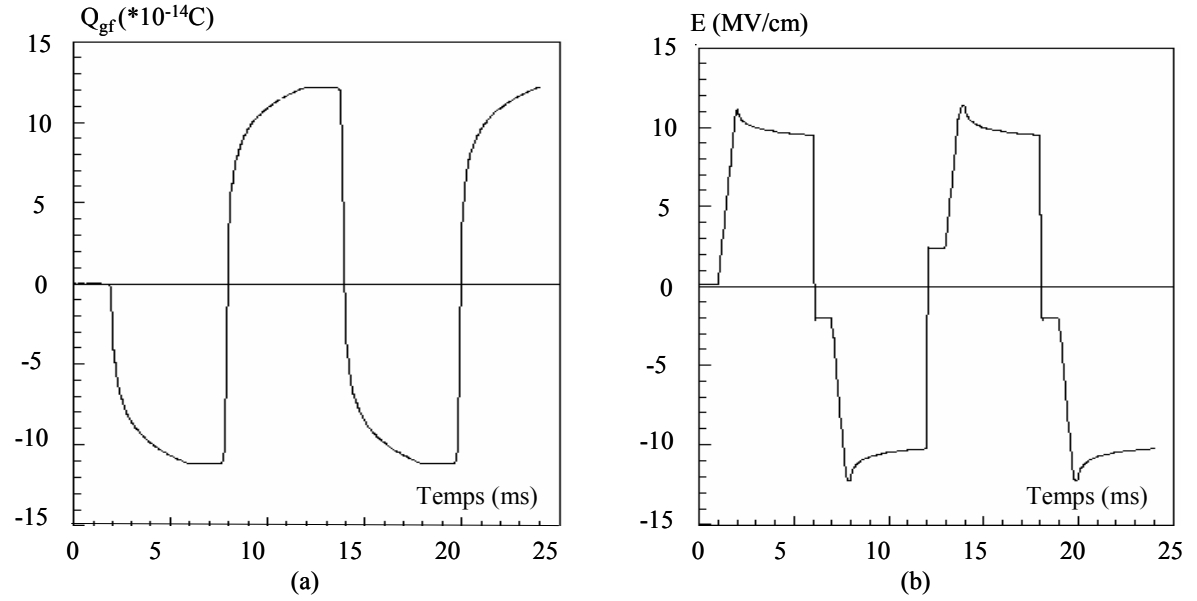


Figure II. 12 Variation de la charge de grille flottante Q_{gf} (a) et du champ électrique E aux bornes de l'oxyde tunnel (b) durant deux cycles de programmation

La figure II.12a représente l'évolution de la charge de grille flottante durant deux cycles de programmation. La mesure de la charge emmagasinée dans la grille flottante peut être réalisée soit après une opération d'écriture soit après une opération d'effacement. Elle représentera la charge Q_{fg0} pour les simulations statiques des cellules écrites et effacées.
Enfin, la figure II.12b montre l'évolution du champ électrique aux bornes de l'oxyde tunnel. Cette variation du champ électrique est corrélée avec l'évolution de la charge de grille flottante et, par conséquent, avec celle du potentiel de grille flottante.

b. Simulations statiques

De manière à mettre en évidence les dépendances géométriques du modèle, des simulations transitoires ont été effectuées avec le changement des paramètres suivants au niveau du transistor mémoire :

- réduction de 50% de la largeur W du transistor mémoire, cette configuration est appelée W_{min},
- réduction de 50% de la longueur L du transistor mémoire, cette configuration est appelée L_{min},
- les simulations effectuées pour les valeurs nominales de ces deux derniers paramètres sont appelées configurations Nom.

A partir de ces simulations transitoires, la valeur de la charge de grille flottante est extraite pour chaque état de la mémoire et pour chaque configuration. Des simulations statiques sont ensuite réalisées pour mettre en évidence l'impact des paramètres géométriques sur les courbes $I_d(V_{gc})$ de la cellule EEPROM. La figure II.13 résume ces différents effets. On peut constater que pour des variations de même ordre de grandeur, l'impact de la largeur du transistor mémoire est prépondérant. En effet, la largeur du transistor mémoire intervient directement dans le calcul des capacités C_{ox} et C_{tun}.

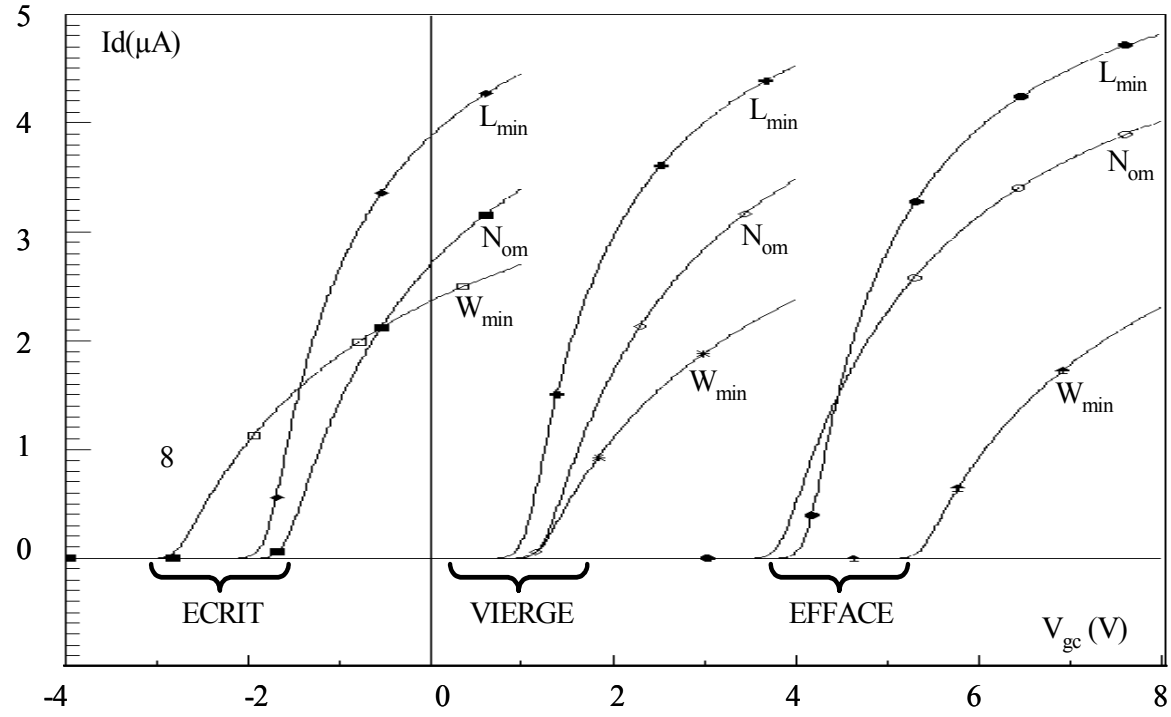


Figure II. 13 Courbes $I_d(V_{gc})$ pour différentes configurations géométriques et pour les trois états de la mémoire.

C. La cellule EEPROM : modèle mathématique

Le diagnostic de défauts géométriques (variation d'un paramètre géométrique de la cellule par rapport aux spécifications) au niveau de la cellule EEPROM nécessite de lier certains paramètres électriques de la cellule mémoire à des paramètres géométriques représentatifs de la cellule (figure II.7).
Pour couvrir tout le domaine de variation des paramètres géométriques, la première approche consisterait à découper l'intervalle de variation de chaque facteur d'entrée. Ainsi une étude qui comporterait sept facteurs avec cent points de variation par facteur, nécessiterait la

réalisation de 100^7 soit 10^{14} expériences de manière à prévoir la ou les réponses correspondantes à n'importe quel point du domaine expérimental défini par les sept facteurs d'entrée.
En fonction du nombre de facteurs, il est souvent impossible de réaliser la totalité des expériences. Il est donc indispensable d'avoir recours à une méthodologie expérimentale ou plan d'expérience.

1 Le plan d'expérience

a. Définition

Une expérience est une épreuve ou un essai effectué pour étudier un phénomène. Dans le domaine scientifique et d'une manière générale, il s'agit d'un fait provoqué ou attendu pour vérifier une hypothèse, une loi, et arriver ainsi à une connaissance théorique de la façon dont se passent les choses. La planification expérimentale propose une série d'outils statistiques pour organiser efficacement des expériences et en analyser les résultats. La plupart de ces méthodes visent à étudier l'effet d'une série de facteurs d'entrée d'un procédé (température, pression, composition d'une formulation...) sur les sorties ou réponses de celui-ci (rendement d'une réaction chimique, viscosité d'un matériau...).
Cette méthodologie permet d'identifier les facteurs expérimentaux les plus significatifs ainsi que les possibles interactions de ces facteurs pour une manipulation donnée. Cependant il peut rester de nombreux facteurs non maîtrisés et donc non contrôlés.
Le concept de plan d'expérience a été introduit au début du siècle dernier (1919) en Angleterre par Ronald A. Fisher dans le secteur agronomique puis, dans l'industrie et dans le domaine biologique et médical. Le passage d'un domaine à l'autre est essentiellement une question de vocabulaire et d'équilibre entre les diverses composantes du plan d'expérience.
Dans le domaine de la microélectronique, les plans d'expériences ou DOE (« Design Of Experiment») sont très efficaces pour améliorer la qualité d'un processus de fabrication par exemple. Ils permettent de déterminer, grâce à une série d'expérimentations judicieusement choisies, les facteurs ou paramètres ayant une influence significative sur la caractéristique à maîtriser. Ainsi, il est possible d'arriver à la connaissance de la réponse pour n'importe quel point du domaine expérimental défini par les facteurs d'entrée. Les différentes étapes qui consistent à maîtriser, décrire, prévoir ou expliquer le phénomène étudié sont :

- la formulation claire du problème étudié,
- l'élaboration de la liste des facteurs susceptibles d'avoir de l'influence, des réponses et des contraintes associées,
- l'établissement d'une stratégie expérimentale ou plan d'expérimentation, c'est-à-dire choisir les expériences que l'on doit effectuer en fonction des objectifs fixés,
- la réalisation des expériences pour obtenir les valeurs des réponses étudiées,
- la déduction des réponses par l'intermédiaire d'un modèle mathématique.

L'étude d'un phénomène commence toujours par l'établissement d'une liste de tous les facteurs qui pourraient avoir une influence sur ce phénomène. Lorsque le nombre de facteurs devient trop important, il convient d'adopter une attitude simplificatrice consistant à réduire le nombre de facteurs étudiés et à les ramener à des valeurs raisonnables. Le poids de chaque facteur est ensuite déterminé expérimentalement lors d'une opération appelée « criblage » où les facteurs les plus influents sur la réponse sont classés pour être réutilisés lors d'une étude ultérieure.

Quand nous connaissons les facteurs les plus influents, nous pouvons, dans une deuxième étape, les étudier de façon plus fine, c'est-à-dire connaître en n'importe quel point du domaine expérimental la valeur d'une ou plusieurs réponses expérimentales. Plus précisément, l'expérimentateur va chercher à prévoir en tout point intérieur au domaine expérimental la valeur de la réponse sans être obligé d'effectuer l'expérience. Pour cela, il faut trouver les relations existant entre les facteurs et les réponses. Cette partie de l'outil méthodologique est appelée « méthodologie des surfaces de réponse ». Cette relation est dans la majorité des cas obtenue en modélisant le phénomène, c'est-à-dire en le simplifiant sous la forme d'un modèle, comme par exemple un modèle mathématique. Ces modèles peuvent être infiniment variés et dépendent du type de problème étudié : modèles linéaires ou non linéaires, équations différentielles…, cependant les modèles les plus utilisés sont les modèles polynomiaux.

b. Application à la cellule EEPROM

Les facteurs d'entrée de notre plan d'expérience sont les sept paramètres géométriques du transistor mémoire EEPROM. Ces paramètres géométriques, présentés figure II.7, ont été choisis de manière à refléter le processus de fabrication des mémoires EEPROM. Il s'agit de :

- L_{eff}, la longueur effective du canal,
- L_{tun}, la longueur de la fenêtre d'oxyde de tunnel,
- $S_{pp}(W_{pp}*L_{pp})$, la superficie de l'oxyde inter polysilicium,
- T_{tun}, l'épaisseur de la fenêtre d'oxyde de tunnel,
- T_{pp}, l'épaisseur de l'oxyde inter polysilicium,
- W_{eff}, la largeur effective du canal,
- T_{ox}, l'épaisseur de l'oxyde de champ au-dessus du canal.

Les trois paramètres électriques choisis comme étant les réponses de nos expériences sont les tensions de seuil de la cellule dans l'état vierge, effacé et écrit ($V_{Tvierge}$, $V_{Tefface}$ et V_{Tecrit}). Ainsi, il sera nécessaire de réaliser une mise en équation de chacune des trois tensions de seuil en fonction des sept paramètres géométriques.
Dans le cadre de notre étude, les expériences seront des séries de simulations effectuées à partir du modèle EEPROM décrit en langage HDL-A. Pour cela, un simulateur SPICE de type ELDO (ANACAD) est utilisé. Cette série de simulations va nous permettre d'avoir une parfaite connaissance des réponses (i.e. tensions de seuil) dans le domaine de variation des sept paramètres géométriques, et cela, à partir d'un nombre limité de simulations.
Le domaine de variation des paramètres géométriques porte le nom de « sphère de connaissance ». Plus précisément, une matrice de Doehlert [DOE78] est utilisée pour définir toutes les configurations des paramètres géométriques à prendre en compte lors des simulations. Cette matrice d'expérience de Doehlert a la caractéristique de présenter une distribution uniforme des points expérimentaux dans l'espace des variables codées. Les points sont disposés suivant un réseau rhombique.
La figure II.14 est un exemple de réseau de Doehlert à trois dimensions (i.e. trois facteurs d'entrée). Ces réseaux uniformes sont particulièrement utiles lorsqu'on veut couvrir un domaine de forme quelconque sans proposer un modèle a priori représentant la réponse.

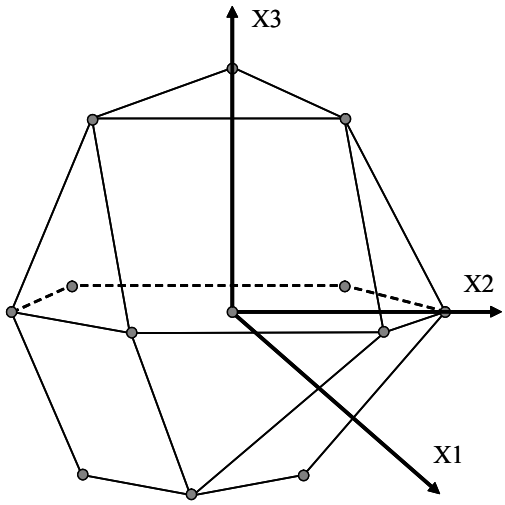


Figure II. 14 Réseau de Doehlert à trois dimensions.

2 Equation polynomiale de la tension de seuil

a. Contexte de simulation.

De manière à générer un modèle polynomial de la tension de seuil du transistor mémoire, des séries de simulations sont réalisées pour des cellules mémoires présentant des géométries différentes. L'étude ne considère qu'une seule cellule, car elle se concentre uniquement sur les paramètres géométriques du transistor à grille flottante. Le but étant bien sûr, de modéliser l'influence des paramètres géométriques du transistor à grille flottante sur les trois tensions de seuil $V_{Tefface}$, V_{Tecrit} et $V_{Tvierge}$ de la cellule mémoire.
Le circuit de simulation est défini sous la forme d'une « netlist » ELDO comprenant le modèle de transistor à grille flottante décrit en HDL-A. De manière à se rapprocher le plus possible des conditions de fonctionnement, les simulations sont effectuées en incorporant dans le circuit tous les éléments nécessaires au fonctionnement de la cellule mémoire EEPROM.
Parmi les différents blocs fonctionnels qui apparaissent sur la figure II.15, on distingue :

- un circuit permettant l'application de la haute tension V_{pp},
- un circuit de décodage de lecture,
- un amplificateur de lecture,
- la cellule mémoire.

Le circuit de routage de la haute tension permet d'appliquer le signal V_{pp} sur les différents nœuds de la cellule mémoire afin de réaliser les opérations de programmation (écriture-effacement). L'application de la haute tension sur le drain du transistor à grille flottante se fait par le biais du transistor SD, alors que l'application de la haute tension V_{pp} sur la grille se fait par le biais du transistor SG. Cette partie logique est commandée par divers signaux désignés génériquement par C_{RL}.
Le décodeur de lecture permet de sélectionner une cellule particulière durant une opération de lecture en fonction des signaux C_{RD} présentés à son entrée. L'amplificateur de lecture produit en sortie une tension V_{out} correspond à la valeur logique de la cellule adressée quand le signal de sélection C_{SA} est actif.
Une capacité de 14 pF, non représentée, est connectée entre la ligne de « bit line » (BL) et la masse pour simuler la charge capacitive induite par les autres cellules d'une colonne du plan

mémoire. Le circuit EEPROM est décrit sur la base des règles de conception de la technologie F6SDP 7.7µm^2 de la société ST-Microelectronic.

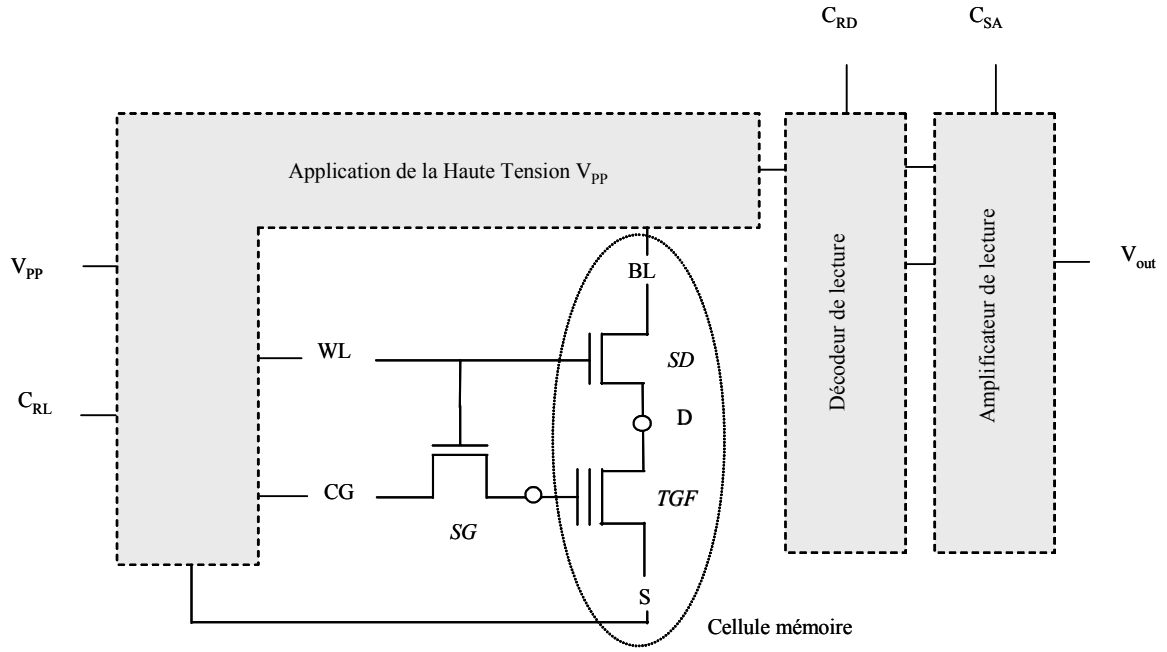


Figure II. 15 Circuit de simulation.

b. Modèle polynomial des tensions de seuil $V_{Tefface}$, V_{Tecrit} et $V_{Tvierge}$.

La figure II.16 résume les différentes étapes qui permettent de générer le modèle mathématique des tensions de seuil.

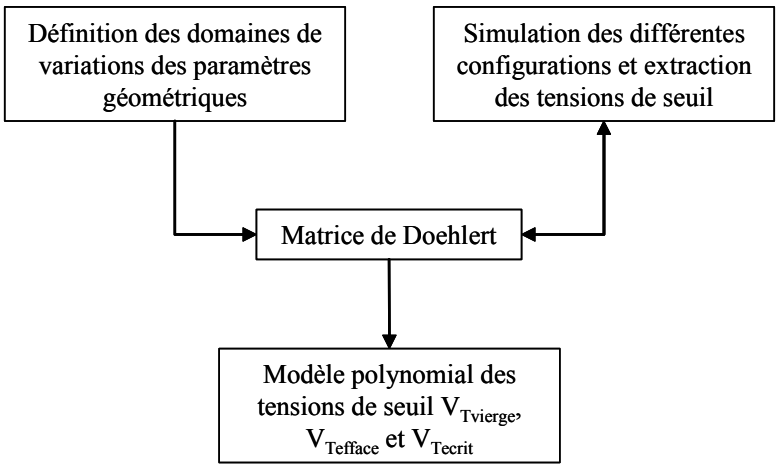


Figure II. 16 Génération du modèle polynomial.

L'équation polynomiale de la tension de seuil est obtenue à partir d'une analyse appelée « surface de réponse ». Chaque paramètre géométrique est associé à un facteur, si bien que sept facteurs sont nécessaires dans le cadre de cette étude. Dans notre exemple d'application, les niveaux choisis pour ces facteurs, conformément à la matrice de Doehlert, sont les suivants :

- L_{eff} est associée au facteur X3 avec 7 niveaux,
- L_{tun} est associée au facteur X5 avec 7 niveaux,
- T_{tun} est associée au facteur X2 avec 7 niveaux,
- S_{pp} est associée au facteur X7 avec 3 niveaux,
- T_{pp} est associée au facteur X1 avec 5 niveaux,
- W_{eff} est associée au facteur X4 avec 7 niveaux,
- T_{ox} est associée au facteur X6 avec 7 niveaux.

A partir de ces informations, une table de simulation est construite, comprenant 57 configurations géométriques différentes. Ces 57 configurations conduisent à la réalisation de 57 simulations électriques sous ELDO. L'extraction des tensions de seuil associées à chaque configuration géométrique permet l'obtention d'un modèle polynomial reliant les trois tensions de seuil ($V_{Tefface}$, $V_{Tvierge}$, V_{Tecrit}) d'une part, aux sept paramètres géométriques du transistor à grille flottante d'autre part. Un algorithme dit de régression multiple, basé uniquement sur les résultats de simulation, est utilisé pour le calcul du modèle polynomial.
Les trois équations polynomiales des tensions de seuil prennent la forme générale suivante :

$$VT = b_0 + \sum_i bi.Xi + \sum_{ii} bii.(Xi.Xj) + \sum_{ij} bij.(Xi.Xj) \qquad \text{(II. 41)}$$

Dans l'équation II.41, les coefficients $b_{i,j}$ sont générés par le plan d'expérience et les variables $X_{i,j}$ sont les paramètres géométriques de la cellule EEPROM.
Le Tableau II.1 représente une partie de la table de simulation (ou plan d'expérience) qui comprend les douze premières configurations à simuler. On distingue d'une part, les valeurs des paramètres géométriques et, d'autre part, les tensions de seuil correspondantes, calculées après simulation. Ce plan d'expérience a été obtenu à partir du logiciel NEMROD [MAT99].

N° Exp	Ttun	Leff	Weff	Ltun	Tox	Spp	VTefface	VTecrit	VTvierge
	A	um	um	um	A	um2	v	v	v
1	85.0	0.7	0.420	0.400	276.0	2.3	3.450	-1.480	1.200
2	85.0	0.7	0.420	0.400	276.0	2.3	3.520	-1.390	1.220
3	98.0	0.7	0.420	0.400	276.0	2.3	3.480	-1.400	1.210
4	72.0	0.7	0.420	0.400	276.0	2.3	2.470	-1.380	1.250
5	72.0	0.7	0.420	0.400	276.0	2.3	2.600	-1.510	1.200
6	98.0	0.7	0.420	0.400	276.0	2.3	2.550	-1.490	1.250
7	89.3	0.9	0.420	0.400	276.0	2.3	2.430	-1.460	1.300
8	80.7	0.5	0.420	0.400	276.0	2.3	2.520	-1.320	1.250
9	80.7	0.5	0.420	0.400	276.0	2.3	2.510	-1.450	1.230
10	93.7	0.5	0.420	0.400	276.0	2.3	2.000	5.000	4.000
11	89.3	0.9	0.420	0.400	276.0	2.3	2.000	5.000	4.000
12	76.3	0.9	0.420	0.400	276.0	2.3	2.000	5.000	4.000
13	89.3	0.9	0.673	0.400	276.0	2.3	2.000	5.000	4.000

Tableau II. 1 Extrait du plan d'expérience.

Le degré de confiance des équations des trois tensions de seuil est donné par l'analyse du résidu, qui correspond à la différence entre les valeurs des tensions de seuil obtenues par les équations et celles obtenues par simulation. Les expériences montrent que l'on peut obtenir

des valeurs maximales de différence de l'ordre de 10 mV pour une cellule vierge, de 180 mV pour une cellule effacée, et de 80 mV pour une cellule écrite. Elles représentent respectivement moins de 2%, 5% et 6% d'erreur par rapport aux valeurs de seuil nominales. De plus, ces valeurs, données à titre indicatif, concernent les zones « frontières », à la limite de la sphère de connaissance.
Ces résultats sont confirmés par le tracé de la droite d'Henry (normal plot), dont un exemple est donné par la figure II.17, pour le cas d'une cellule vierge. Sur cette figure, les valeurs de résidus négatives et positives sont présentées en abscisse, avec la probabilité liée à ces valeurs en ordonnée.

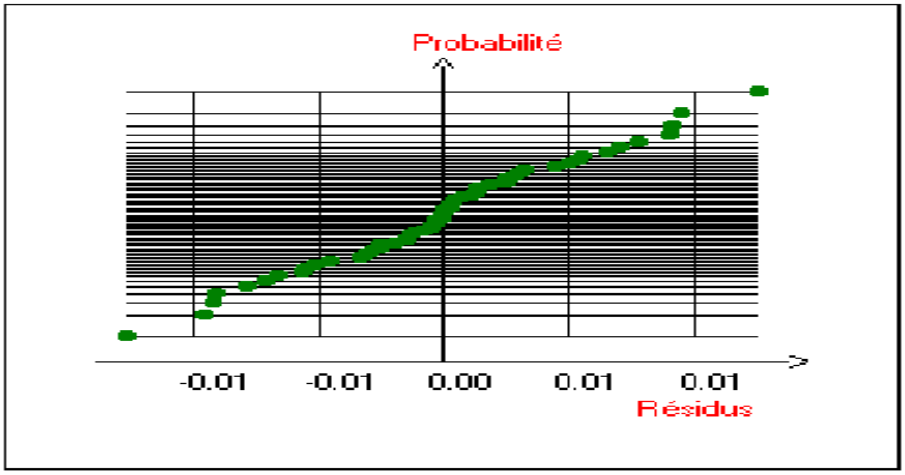


Figure II. 17 Calcul des résidus (logiciel NEMROD).

Si les résidus suivent une distribution normale, les points sont presque alignés. Si des points s'écartent de la ligne, ils doivent être vérifiés. Une disposition particulière des points peut indiquer la nécessité d'une transformation des données pour rendre leur distribution normale.

Les principaux avantages offerts par l'utilisation d'un modèle polynomial des tensions de seuil sont :

- une expression polynomiale simple qui procure des temps de calcul courts, permettant ainsi d'envisager de nombreuses variations de géométries,
- un nombre relativement restreint de simulations (57 dans notre cas), permettant d'obtenir un temps de génération du modèle mathématique court. Cette qualité peut être exploitée pour étalonner rapidement et automatiquement le modèle lors d'un passage d'une technologie à une autre, ou d'une unité de production à une autre.

III. Diagnostic des défauts géométriques dans l'EEPROM

A. Méthodologie de diagnostic

La méthodologie de diagnostic de défauts géométriques dans l'EEPROM commence par l'extraction des tensions de seuil. Ensuite, à partir du modèle mathématique de la tension de seuil, toutes les géométries (groupes constitués des sept paramètres géométriques définissant la cellule) qui correspondent aux tensions de seuil extraites sont générées. Enfin, une étude statistique des géométries générées permet de dresser les probabilités de défaillance pour chaque paramètre géométrique.

1 Extraction des tensions de seuil

L'extraction des trois tensions de seuil se fait lors des étapes de caractérisation électrique de la cellule mémoire. Cette caractérisation est réalisée sur des structures de test (TEG) qui se trouvent dans les lignes de découpe des plaquettes de silicium. Ces structures permettent la caractérisation de cellules EEPROM isolées, ce qui n'est pas le cas pour les cellules EEPROM des puces mémoires. Nous avons choisi, pour notre étude, des structures de test comprenant le transistor de sélection ainsi que le transistor à grille flottante.
La programmation de la cellule mémoire, ainsi que la mesure des tensions de seuil se fait dans les mêmes conditions que celles utilisées lors des simulations. Les conditions de programmation (niveaux de tension et formes d'onde) utilisées lors de la programmation des cellules sur structures de test sont données figure II.18.

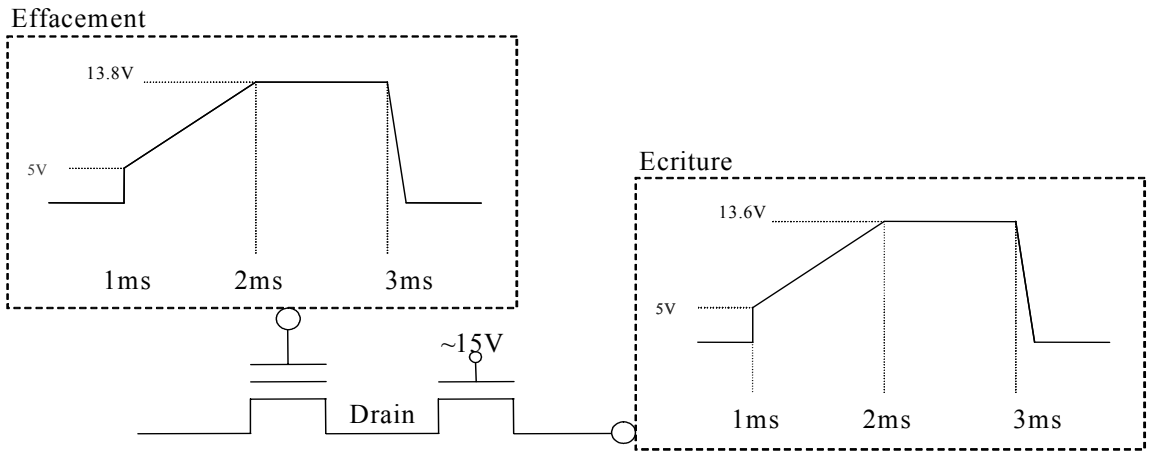


Figure II. 18 Conditions de programmation.

La mesure des tensions de seuil passe par l'extraction des caractéristiques $I_d(V_{gc})$ des cellules mémoires dans leur état électrique vierge, écrit et effacé. La tension de seuil est ensuite calculée à partir des courbes de transconductance.
Les trois tensions de seuil correspondant à une cellule mémoire spécifique serviront d'entrée à un module de calcul dynamique. Ce module de calcul va générer une liste de toutes les géométries correspondant aux tensions de seuil d'entrée.

2 Générations des géométries

Le principe de génération des géométries candidates est illustré figure II.19. Les trois tensions de seuil obtenues à partir de la caractérisation électrique d'une cellule mémoire sont définies au sein d'une fenêtre de variation (correspondant en général aux incertitudes sur la mesure et sur l'extraction des tensions de seuil). A partir de cette fenêtre de variation ΔV_T des tensions de seuil, des géométries dites « candidates » sont générées à l'aide d'un module de calcul. Ce module de calcul utilise le modèle mathématique de la tension de seuil et génère de manière dynamique les géométries candidates.
L'exploitation du modèle mathématique de la tension de seuil se fait au sein d'un programme informatique développé en langage C.
La liste des géométries candidates représente toutes les configurations géométriques correspondant aux tensions de seuil cibles. Le nombre des géométries candidates dépend d'une part de la fenêtre de variation des tensions de seuil initiales et d'autre part, du pas de calcul utilisé pour chaque paramètre géométrique. Cette liste des géométries candidates sera traitée lors d'une phase d'exploitation des données générées par l'outil informatique. Cette

étude fournira la probabilité de défaillance associée à chaque paramètre géométrique (i.e. paramètres hors spécifications).

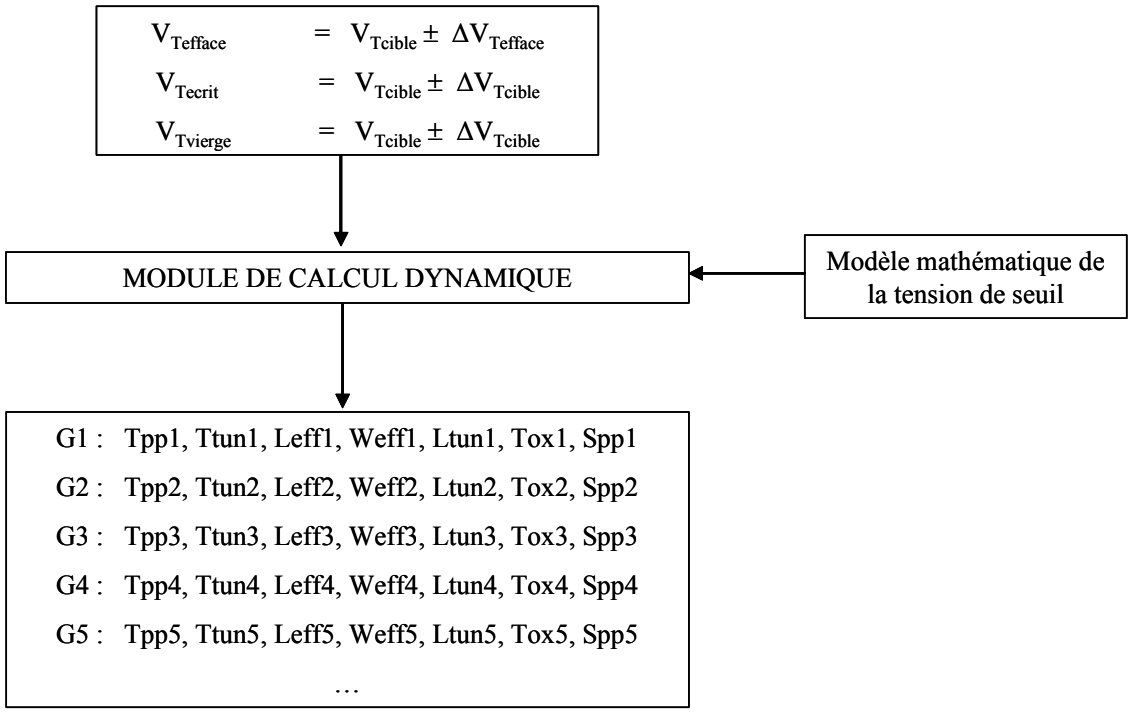


Figure II. 19 Génération des géométries candidates

3 Probabilité de défaillance de chaque paramètre géométrique

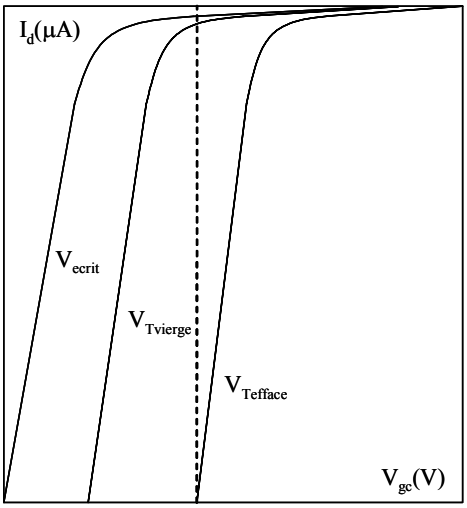
La dernière étape de la méthodologie de diagnostic des défauts géométriques dans la cellule EEPROM correspond à l'exploitation des résultats obtenus lors de l'étape de génération des géométries candidates. Une étude statistique des géométries candidates permet de mettre en évidence le ou les paramètres géométriques défaillants.
Pour établir ces probabilités de défaillance, les valeurs des paramètres géométriques générés par l'outil sont comparées d'une part aux valeurs nominales de chaque paramètre géométrique et d'autre part, aux tolérances de variations associées à chaque paramètre.
Ces tolérances sont obtenues en considérant la technologie de fabrication utilisée, et plus particulièrement le processus de fabrication associé à l'étape qui définit chaque paramètre géométrique.
Le temps de génération et de traitement varie de quelques secondes à quelques minutes en fonction de la fenêtre de variation des tensions de seuil cibles, de la fenêtre de variation de chaque paramètre géométrique et du pas de calcul utilisé pour chaque paramètre géométrique.

B. Validation silicium et outil logiciel

1 Validation silicium de la méthodologie de diagnostic

a. Extraction des tensions de seuil d'une cellule isolée

La première étape de validation sur silicium de l'outil de diagnostic consiste à extraire les trois tensions de seuil relatives à une cellule EEPROM.



Mesure des tensions cibles :

V_{Th}efface = 4.2 ± 0.01 V

V_{Th}ecrit = -2.2 ± 0.01 V

V_{Th}vierge = 1.2 ± 0.01 V

Filtre : 0.01V pour chaque état

Figure II. 20 Extraction des tensions de seuil.

Ces trois tensions cibles ont été choisies de manière à présenter de légères variations par rapport aux valeurs standard de la technologie. La fenêtre de variation de ces tensions de seuil (ou « filtre » de l'outil de génération des géométries) est fixée à 0.01V comme présenté figure II.20.

b. Génération de géométries candidates

A partir des tensions de seuil cibles, l'outil de diagnostic génère une liste de géométries candidates dont une partie est présentée dans le tableau II.2.
Les géométries candidates apparaissent à gauche du tableau avec les tensions de seuil correspondantes sur la droite.

Tpp,	Ttun,	Leff,	Weff,	Ltun,	Tox,	Spp	V_{th}efface	V_{th}ecrit	V_{th}vierge
200.5	79.0	0.705	0.49	0.55	266.0	3.055	4.19	-2.19	1.19
226.0	79.0	0.669	0.43	0.43	288.0	2.480	4.20	-2.19	1.20
251.5	79.0	0.634	0.40	0.46	266.0	2.710	4.19	-2.19	1.20
115.5	82.0	0.882	0.43	0.52	266.0	2.135	4.19	-2.20	1.20
209.0	82.0	0.918	0.40	0.40	299.0	2.595	4.20	-2.19	1.19
183.5	85.0	0.811	0.25	0.43	200.0	2.480	4.19	-2.20	1.20
….....									

Tableau II. 2 Génération des géométries candidates.

c. Analyse des résultats et validation silicium

Après traitement statistique, les probabilités de défaillance associées à chaque paramètre géométrique mettent en évidence les paramètres les plus défectueux. L'exemple de la figure II.21a fournit une classification des probabilités de défaillance de chaque paramètre géométrique. Le paramètre dont la probabilité de défaillance est la plus importante est la largeur effective du canal W_{eff} avec une probabilité de défaillance P_{weff} de 100%.
Cette probabilité de défaillance est confirmée par le tracé la distribution de toutes les valeurs du paramètre W_{eff} (figure II.21b). Cette distribution montre une population de valeurs du paramètre géométrique W_{eff} en dehors de sa fenêtre de spécification.

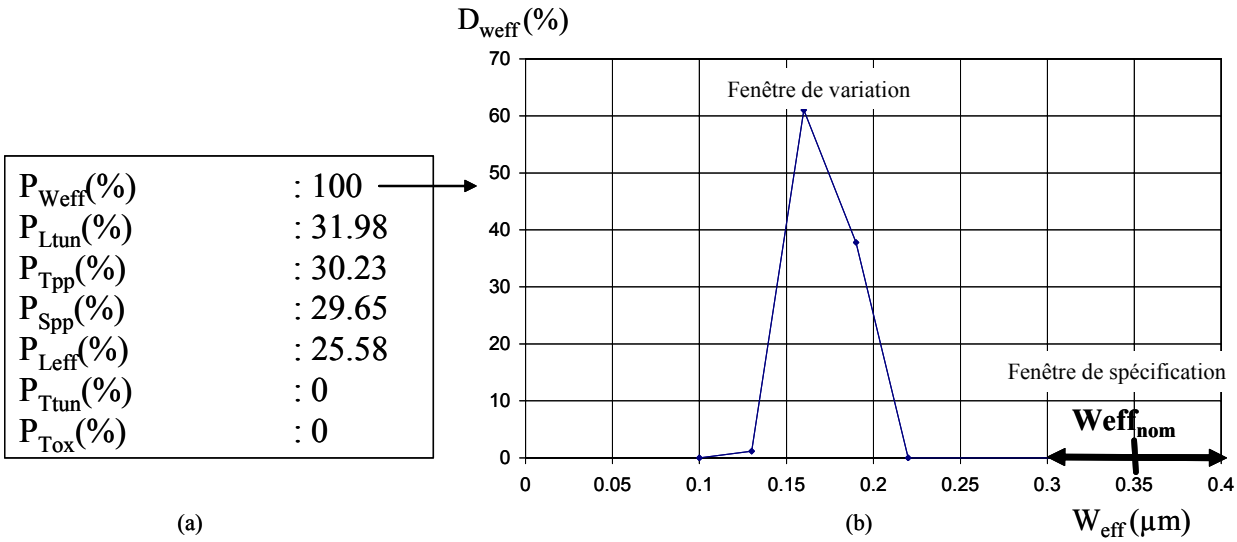


Figure II. 21 Probabilités de défaillance (a) et distribution du paramètre le plus défectueux (b).

De même que pour le paramètre W_{eff}, la distribution des valeurs de tous les autres paramètres géométriques peut être obtenue.
De manière à valider les résultats obtenus, une coupe transversale, effectuée sur le transistor à grille flottante sur lequel ont été extraites les tensions de seuil cibles, met en évidence une croissance de l'oxyde de champ anormalement importante, qui a pour effet de réduire considérablement la largeur effective du canal du transistor (figure II.22).

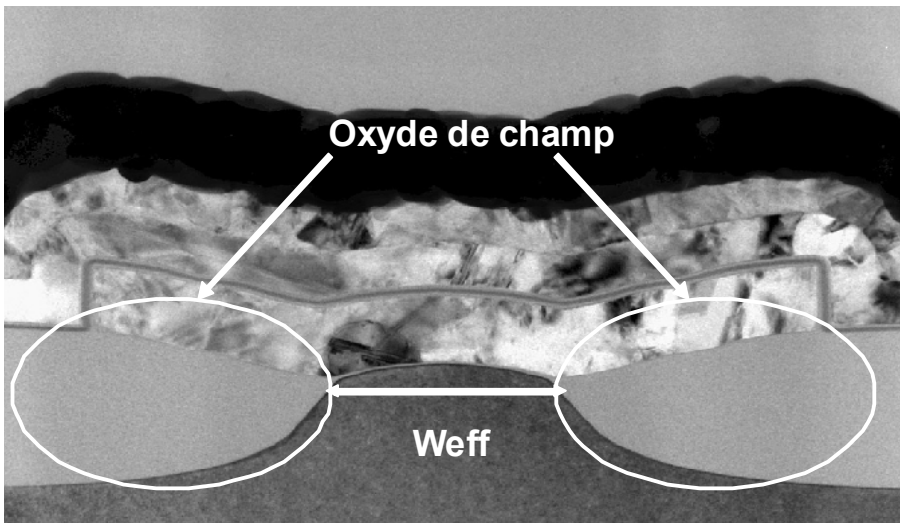


Figure II. 22 Largeur effective du canal W_{eff} sous dimensionnée

Cet exemple de validation de notre méthodologie de diagnostic de défauts géométriques a été réalisé sur une cellule mémoire fabriquée suivant une technologie « F6DP7.7µm^2 » (STMicroelectronics). Cette méthodologie montre de façon claire que le processus de diagnostic de défauts géométriques dans la cellule EEPROM peut être rapidement réalisé à partir de l'extraction de signatures électriques représentatives d'une cellule mémoire EEPROM, comme les tensions de seuil.

2 Outil logiciel

De manière à automatiser la méthodologie de diagnostic de défauts et dans un souci de convivialité, nous avons développé un outil logiciel basé sur une interface graphique. Les principales entrées relatives à l'outil sont :

- les valeurs de tension de seuil cibles (correspondant à une cellule dans l'état électrique vierge (UV), effacée et écrite),
- la précision ou filtre associée à chacune des trois tensions de seuil,
- les limites inférieures et supérieures tolérées de chaque paramètre géométrique,
- le pas de découpage associé à chaque paramètre géométrique (nombre de points).

Les deux principales fenêtres d'entrées de l'outil logiciel sont présentées sur la figure II.23. Comme nous l'avons vu, en fonction des entrées du logiciel, les résultats fournis après traitement sont de deux types :

- la probabilité de défaillance associée à chaque paramètre géométrique,
- la distribution de chaque paramètre géométrique.

La figure II.24 donne les probabilités de défaillance de paramètres géométriques suivantes :

- W_{eff} possède une probabilité de défaillance de 62%,
- L_{eff} possède une probabilité de défaillance de 12%,
- L_{tun} possède une probabilité de défaillance de 50%,
- T_{tun} possède une probabilité de défaillance de 0%,
- T_{ono} (T_{pp}) possède une probabilité de défaillance 37%,
- T_{ox} possède une probabilité de défaillance de 50%,
- S_{ono} (S_{pp}) possède une probabilité de défaillance de 75%.

Ces probabilités de défaillance ont été générées pour les tensions de seuil d'entrées qui apparaissent figure II.23 (tensions de seuil données uniquement à titre d'exemple) :

V_{Tecrit} = -1.49 V
$V_{Tvierge}$ = 1.18 V
$V_{Tefface}$ = 3.44 V

En plus de ces premières informations, le tracé de la distribution de chaque paramètre géométrique peut être obtenu. L'exemple de la distribution du paramètre L_{eff}, présentée figure II.24, met en évidence une queue de distribution hors des limites de spécification.

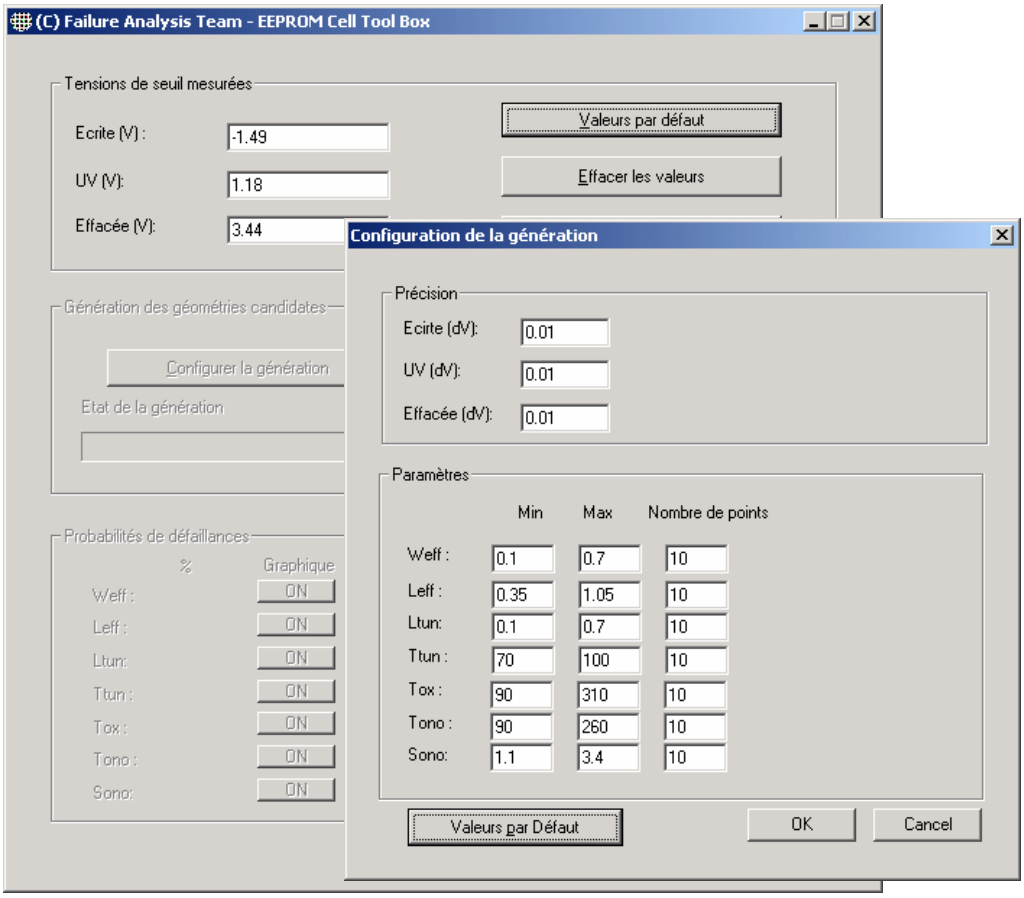


Figure II. 23 Entrées de l'outil de diagnostic.

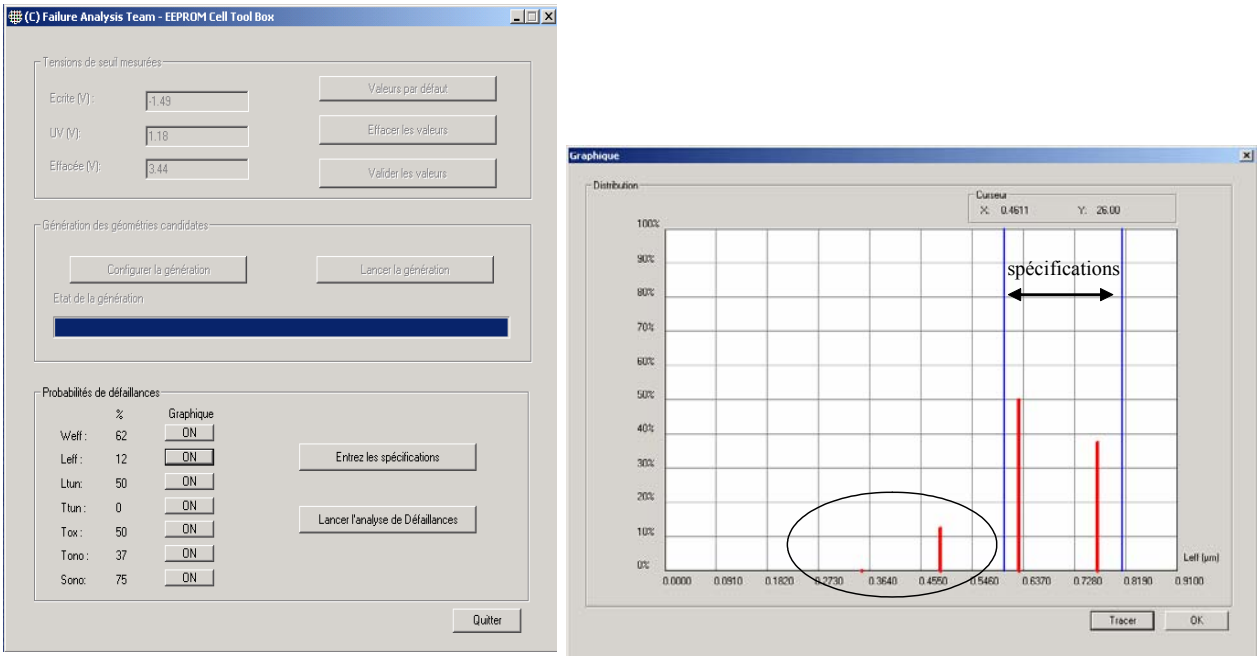


Figure II. 24 Résultats de sortie de l'outil de diagnostic.

C. Conclusion

Une très faible variation des paramètres géométriques d'une cellule EEPROM peut entraîner une variation importante au niveau de ses tensions de seuil, et donc un impact au niveau des performances globales de la cellule. De ce fait, une méthodologie de diagnostic de défauts géométriques rapide et fiable, basée uniquement sur l'extraction des tensions de seuil d'une cellule isolée a été mise en place. Cet outil s'est avéré efficace pour déterminer l'étape de fabrication (liée au paramètre géométrique défectueux) responsable d'une défaillance.
La connaissance des paramètres géométriques défectueux peut permettre par la suite de corriger l'étape du processus de fabrication responsable de l'obtention de signatures électriques erronées ou éloignées des spécifications.

CHAPITRE III
Diagnostic de Défauts dans le plan Mémoire EEPROM

I. Analyse de défaillance dans le plan mémoire EEPROM

A. Architecture des mémoires EEPROM

1 Le plan mémoire EEPROM

Les caractéristiques principales des mémoires EEPROM sont :

- une granularité minimale puisque l'effacement peut s'effectuer sur un mot particulier du plan mémoire,
- une très bonne tenue en endurance (jusqu'à 1 million de cycles),
- des temps de programmation longs (entre 2 ms et 10 ms),
- des capacités de stockage pouvant aller jusqu'à 1Mbit,
- une interface de communication série et parallèle.

La figure III.1 montre la place occupée par les EEPROM parmi les mémoires non volatiles. Le principal inconvénient des EEPROM est la taille du point mémoire qui est quatre fois plus élevée que celle d'une cellule mémoire de type Flash par exemple. Ce qui entraîne une surface occupée par le plan mémoire EEPROM importante (pour une capacité donnée) et un coût par bit élevé.

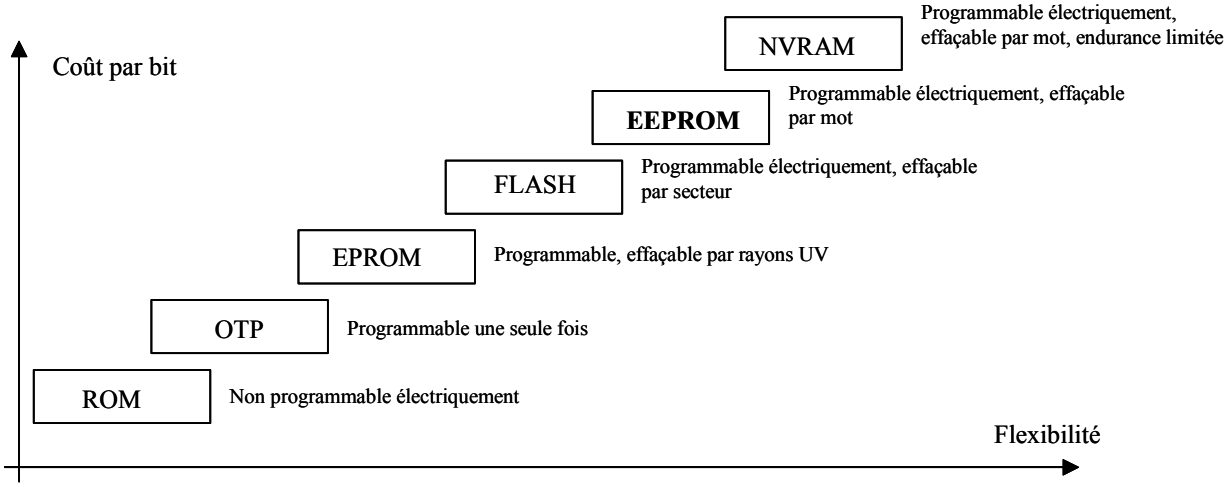


Figure III. 1 Les mémoires EEPROM parmi les mémoires non volatiles.

2 Les différents types de mémoires EEPROM

a. Mémoires EEPROM à accès série et à accès parallèle

Les mémoires EEPROM peuvent se présenter sous une forme isolée (stand alone) ou embarquées dans des produits spécifiques comme les cartes à puces.
Au sein de la famille des mémoires isolées, on retrouve les mémoires à accès parallèle et les mémoires EEPROM pourvues d'interfaces de communication séries.
Une mémoire EEPROM à accès parallèle est contenue dans un composant comportant 28 à 32 broches, alors qu'un composant disposant d'un protocole de communication série ne comporte en général que 8 broches externes. L'utilisation des mémoires à accès série permet donc de réduire considérablement le coût du produit final encapsulé. De plus, l'interface de

communication série facilite l'intégration de ce type de mémoire dans des systèmes à microprocesseurs.
Néanmoins, les mémoires à accès parallèle offrent des temps d'accès rapides (100 ns), contrairement aux mémoires à accès série où l'accès à une donnée en mémoire passe par la transmission des signaux d'instructions et d'adresse en série (24 µs à 30 µs suivant la fréquence d'horloge).
Concernant les interfaces de communication série, les plus utilisées sont les interfaces de communication I^2C, SPI et MICROWIRE, dont une brève description va suivre.

Le bus I^2C (Inter Integrated Circuit) a été développé au début des années 80 par la société Philips semiconductors pour permettre de relier facilement à un microprocesseur les différents circuits d'un téléviseur moderne. Les composants répondant aux spécifications I^2C sont de plus en plus utilisés. En effet, le fait d'avoir un bus ne comportant que deux fils pour accéder à ce type de composant présente un avantage évident. Il est aisé de les intégrer dans un système à microcontrôleur, par exemple, car ces composants n'occupent que peu de place et utilisent un nombre limité de broches. Cependant, les données sont transmises en série à 100 MHz en mode standard et jusqu'à 400 MHz en mode rapide, ce qui ouvre la porte de cette technologie à toutes les applications où la vitesse n'est pas primordiale [PHI00].

Le fonctionnement des mémoires SPI (Serial Peripheral Interface) est basé sur un échange d'informations avec l'extérieur sur un fil d'entrée et un fil de sortie, avec une sélection (autorisation d'accès) de la mémoire. Le bus SPI est une liaison série synchrone qui opère en mode « full duplex » (émission et réception simultanées). La gamme des produits SPI est un ensemble de mémoires EEPROM, d'une capacité allant de 1 Kbit à 64 Kbit, dédié principalement aux applications à microcontrôleurs. Les mémoires utilisant ce bus, élaboré par MOTOROLA pour ses propres microcontrôleurs, sont aujourd'hui proposées par de nombreux fabricants. Avec ce protocole de communication, la sélection de la mémoire est réalisée par une voie indépendante contrairement au protocole I^2C où chaque composant possède une adresse propre. L'espace d'adressage mémoire est ainsi illimité. La vitesse de fonctionnement de ce type de circuit est supérieure à 2 MHz avec une très bonne immunité au bruit. Le marché des mémoires SPI est encore un marché limité, tourné essentiellement vers les applications automobiles où la vitesse de circulation de l'information est primordiale (paramètre de contrôle de l'injection, freinage ABS, ….).

L'interface MICROWIRE est la première interface série à avoir été utilisée, elle est par conséquent largement répandue. Cependant son architecture est dépassée et les nouveaux circuits utilisent plutôt des interfaces de type SPI ou I^2C.

b. Mémoire EEPROM embarquée : la carte à puce

L'évolution du marché des dispositifs portatifs tels que des cartes à puce ou des supports de communication mobiles, influe directement sur l'utilisation des mémoires EEPROM (mémoires programmables et effaçables électriquement). En effet, ces dernières sont devenues au cours des dernières années la solution pour toute application faisant appel à une mémoire semi-conducteur non-volatile.
La carte à puce est avant tout un composant réunissant sur une même puce un cœur de processeur optimisé pour gérer des interruptions, ainsi que des mémoires et d'autres périphériques utiles à la réalisation d'applications de contrôle et de sécurité.
La taille limite de la puce encartée (25 mm^2) influence les choix d'évolution pris par les fabricants de cartes à puce. Par exemple, les concepteurs cherchent toujours à augmenter la quantité de mémoire embarquée (ROM, RAM et EEPROM) ainsi que leur qualité (avènement des mémoires Flash et Fram dans certains microcontrôleurs encartables), leur but principal étant d'approcher le meilleur compromis entre encombrement, vitesse d'accès, consommation

et endurance. Les tendances de ces dernières années s'orientent vers un cœur de processeur puissant, une unité d'exécution pour le cryptage, de grandes quantités de mémoire et une double interface de communication (avec contact et sans contact).
La figure III.2 montre un exemple d'architecture de produit carte à puce embarquant de la mémoire EEPROM. Pour des raisons de confidentialité, nous n'entrerons pas dans les détails de fonctionnement de ce type de produit fabriqué par la société STMicroelectronics.
On distingue trois types de mémoires embarquées : de la mémoire vive de type SRAM, de la mémoire ROM et de la mémoire non volatile reprogrammable de type EEPROM. La non volatilité de ces dernières permet de stocker des informations dans des applications portables. Le test de la carte à puce passe par le test de chacun de ses éléments et conduit à la réalisation de flots de tests denses et complexes. La principale cause de perte de rendement de ce type de produit est liée à la partie EEPROM.
En effet, comme le montre la figure III.3, la partie EEPROM occupe environ 50% de la surface du produit, ce qui rend l'EEPROM embarquée particulièrement sensible à tous types de défaillances affectant la puce (particules aléatoires, variation du processus de fabrication…).
De plus, la structure et le principe de fonctionnement spécifique de la mémoire EEPROM (oxydes fins, tensions de programmation élevées, densité du plan mémoire élevée) en font un composant particulièrement sensible, contrairement aux mémoires de type ROM et RAM réputées pour être plus robustes.

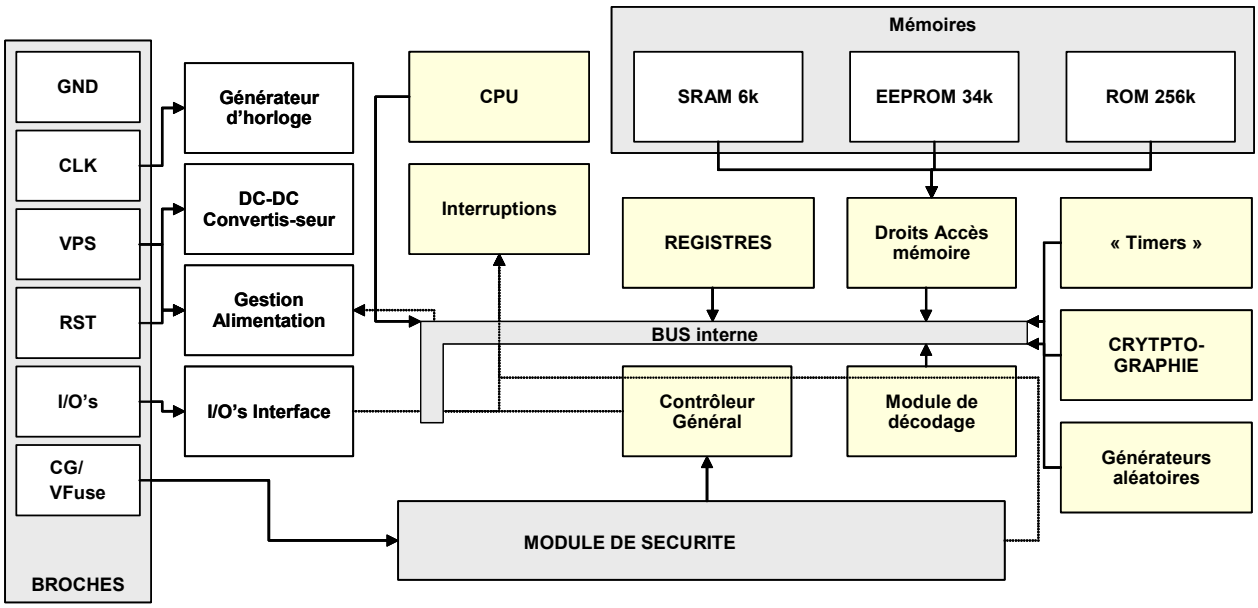


Figure III. 2 Architecture d'un produit carte à puce.

La surface occupée par la partie EEPROM s'explique par la taille importante de la cellule mémoire EEPROM, composée de deux transistors : le transistor mémoire qui assure la fonction de mémorisation et un transistor de sélection haute tension. Les cellules mémoires utilisées actuellement dans les cartes à puces occupent une surface de 3,9 μm^2. A titre de comparaison, une cellule Flash occupe une surface d'environ 0,35 μm^2. Les cellules SRAM et ROM occupent quant à elles respectivement des surfaces de 5 μm^2 et 0,5 μm^2.

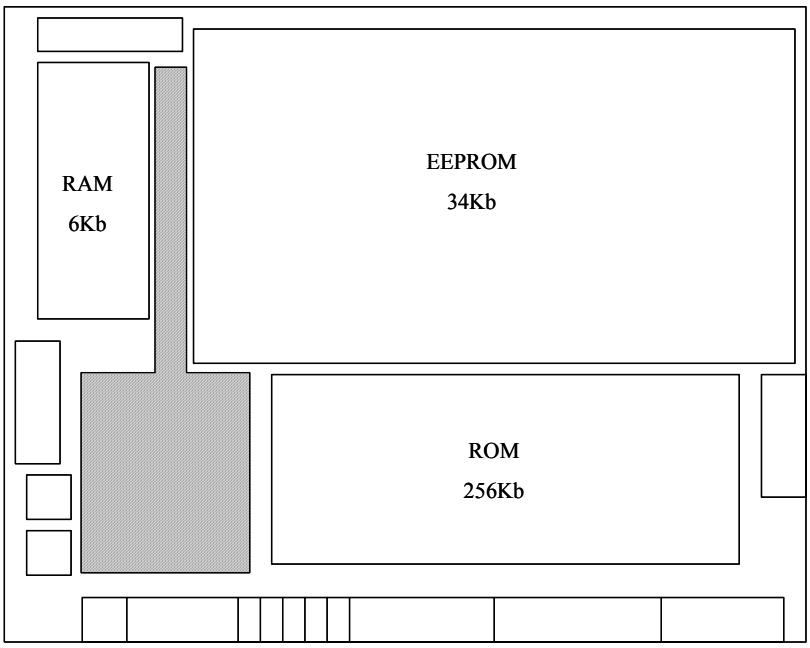


Figure III. 3 Mémoires contenues dans une carte à puce.

B. Technique classique d'analyse de défaillance dans le plan mémoire EEPROM

Un procédé d'analyse de défaillance classique, qui utilise les propriétés spécifiques de régularité du plan mémoire est présenté dans cette partie. Ce procédé d'analyse comprend quatre étapes principales :

- le test des circuits mémoires qui permet d'obtenir des informations quantitatives comme le rendement ou la classe de rejet associée à chaque puce défaillante. Ces tests peuvent s'effectuer soit sur des produits embarquant de la mémoire EEPROM (en utilisant des programmes de test standards), soit à partir de véhicules de test qui permettent une plus grande flexibilité en terme de caractérisation du plan mémoire;
- la localisation des bits (ou groupes de bits) défaillants au sein du plan mémoire de la puce défaillante à partir d'outils d'analyse comme le « Bitmap »;
- la caractérisation électrique des cellules défaillantes à partir de programmes de test spécifiques (on utilise à ce niveau des véhicules de test qui permettent l'accès à des cellules isolées du plan mémoire);
- l'analyse physique des cellules défaillantes.

L'analyse de construction est une technique d'analyse de défaillance classique qui regroupe les étapes de localisation et de caractérisation électrique et physique des cellules défaillantes du plan mémoire.

1 Test de produits embarquant de la mémoire non volatile de type EEPROM

A l'issue du test de produits embarquant de la mémoire EEPROM, deux principaux types d'informations sont donnés. Tout d'abord les informations de « binning » associées à chaque puce testée. Ces informations correspondent à un type de rejet, obtenu à une certaine étape du programme de test. En général, le « bin 0 » correspond toujours à une puce qui n'a subi aucun rejet durant l'exécution du programme de test, c'est-à-dire à une puce bonne. La deuxième information est le rendement de chaque tranche de silicium qui correspond au

rapport entre le nombre de puces bonnes à l'issu du test sur le nombre de puces potentielles que contient la tranche de silicium.
Le test standard est effectué de manière systématique sur toutes les plaques et permet de localiser les puces défaillantes. La figure III.4 est un exemple de cartographie de classe de rejets obtenu pour chaque puce testée sur la tranche de silicium. On peut noter que chaque puce est associée à une couleur spécifique et à un numéro (au centre de la puce) qui correspond à la classe de rejet.

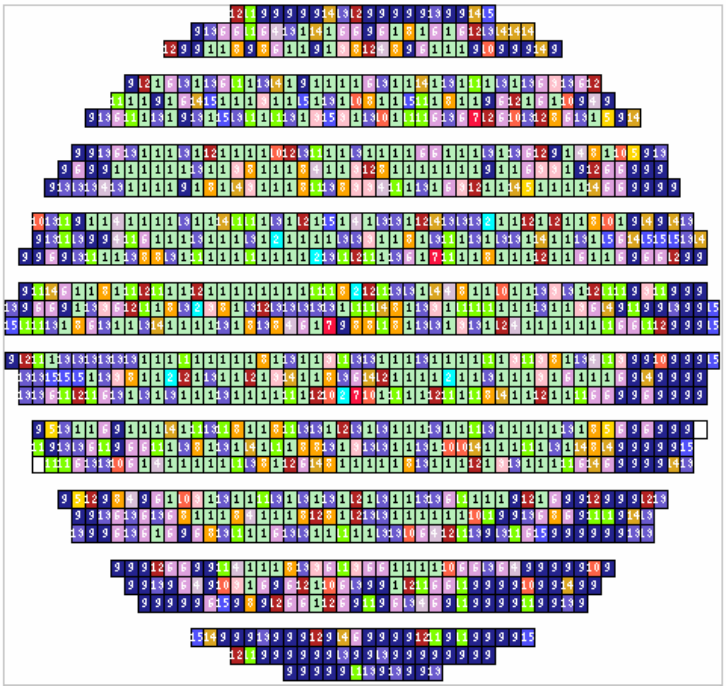


Figure III. 4 « Binning » obtenu à l'issu d'un flot de test standard.

2 Test des mémoires EEPROM isolées : le véhicule de test EEPROM 0.18 µm

Dans le cas des mémoires EEPROM embarquées, l'accès et l'analyse des défauts affectant le plan mémoire est très difficile. Une alternative à cette limitation a été d'introduire des véhicules de test qui accompagnent les produits embarquant la mémoire EEPROM. L'architecture de ces véhicules de test se rapproche des mémoires à accès parallèle. Ils permettent une caractérisation précise du plan mémoire (accès à certains signaux internes au plan mémoire, modes de test spécifiques, accès aux cellules élémentaires du plan mémoire,…).
Le véhicule de test EEPROM qui accompagne l'introduction des technologies cartes à puce avancées se compose d'une matrice de cellules 512 Kbits avec uniquement la circuiterie de décodage et de lecture. Tous les signaux de programmation sont fournis extérieurement, ce qui permet une analyse plus fine des causes de rejets. L'architecture de ce véhicule de test sera détaillée au Chapitre IV.
La figure III.5 présente des résultats de tests obtenus sur le véhicule EEPROM 512 Kbits pour 8 tranches de silicium. On distingue clairement les classes de rejet pour chaque puce testée. En plus de ces informations quantitatives, la spécificité de ce véhicule de test permet d'obtenir des informations cette fois qualitatives comme les distributions des tensions de seuil (figure III.7) ou encore les courbes de transfert d'une cellule isolée du plan mémoire. Ces dernières informations permettent, par exemple, de caractériser de manière plus précise une cellule défaillante du plan mémoire.

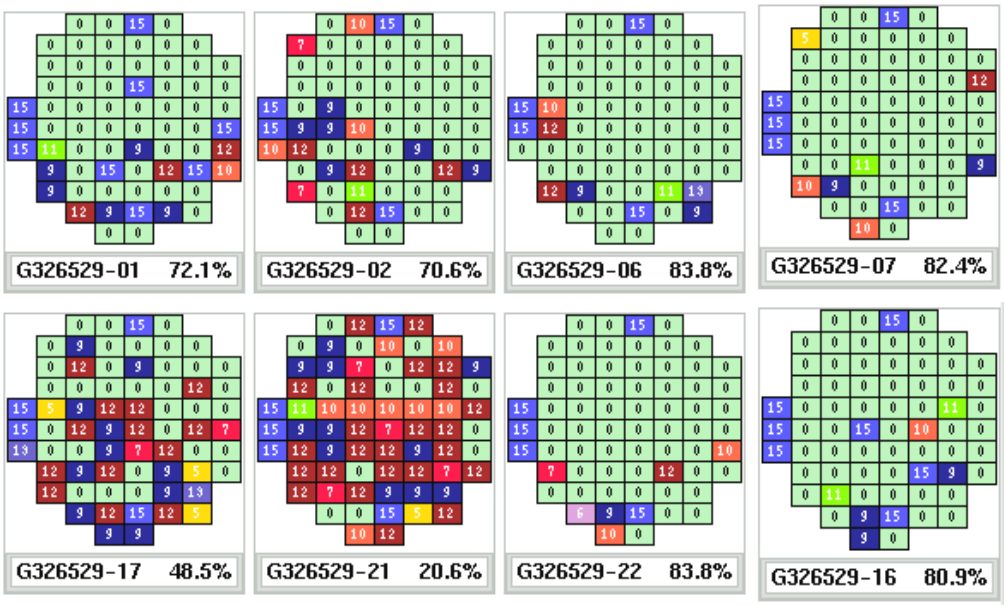


Figure III. 5 « Binning » obtenu à l'issu du test standard du véhicule EEPROM 512 Kbits.

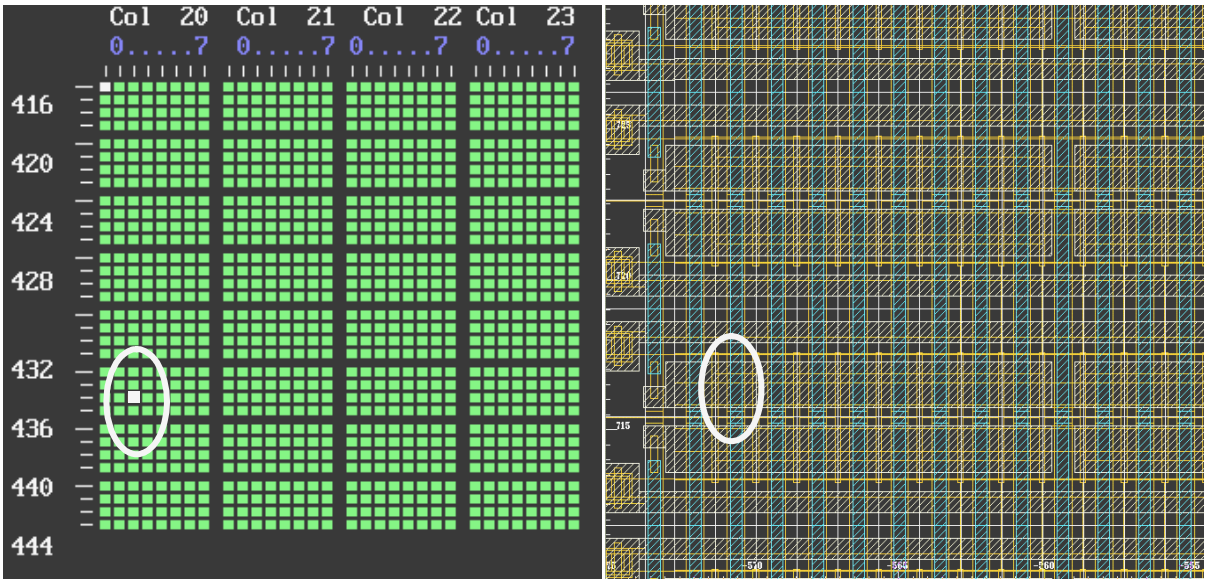
3 L'analyse de construction

a. Présentation

Un exemple d'analyse de défaillance classique, réalisé à partir du véhicule de test EEPROM fabriquée par la société STMicroelectronics (technologie 0.18 µm), est présenté dans cette section. Cette analyse a été effectuée suite à un nombre important de rejets obtenus après les tests de rétention de produits embarquant de la mémoire EEPROM.
L'analyse de construction permet, dans certains cas, de remonter à la cause exacte d'une défaillance affectant le plan mémoire. L'inconvénient principal de cette technique réside dans le temps d'analyse et le coût non négligeable des moyens mis en œuvre.
Je rappelle que ce type d'analyse débute par la localisation des cellules défaillantes du plan mémoire, ensuite des informations électriques relatives aux cellules défaillantes sont obtenues à partir de programmes de caractérisation spécifiques. Enfin, une caractérisation physique est effectuée de manière à tenter de remonter à l'origine physique de la défaillance.

b. Localisation des bits défaillants dans le plan mémoire EEPROM

Une cartographie bit du plan mémoire est généralement disponible à l'issue du flot de test. Cette cartographie est systématiquement sauvegardée en mémoire du testeur lors de rejets obtenus à l'issue des tests fonctionnels.
La figure III.6a est une cartographie bit logique obtenue à l'issue d'un rejet aux tests fonctionnels du véhicule EEPROM 0.18 µm.
Cette cartographie est une représentation topologique des cellules défaillantes du plan mémoire. Elle permet de localiser une cellule défaillante (cellule cerclée) à partir de sa ligne, de sa colonne et de sa position au sein du mot mémoire, constitué ici de huit bits.
Ces informations permettent par la suite de localiser la cellule mémoire défaillante au niveau du dessin des masques ou « layout » du circuit mémoire comme le montre la figure III.6b. Ainsi, il est possible de connaître la position exacte de la cellule défaillante sur la plaquette de silicium.



(a) (b)

Figure III. 6 Cartographies bit logique (a) et du dessin de masques (b).

c. Caractérisation électrique

La caractérisation électrique d'une cellule ou d'un groupe de cellules défaillantes est difficile à mettre en place sur produit. Elle ne peut se faire que si le produit dispose de modes de test spécifiques (DMA, CDMA…) qui permettent un accès sélectif aux cellules du plan mémoire (cf. Chapitre I §III.B.1).
Les véhicules de test qui accompagnent l'introduction de nouveaux produits permettent l'extraction des courbes de transfert $I_d(V_{gc})$ ou $I_d(V_{ds})$ d'une cellule isolée du plan mémoire, mais aussi l'extraction des distributions de tensions de seuil V_{th} (obtenues à un niveau de courant paramétrable). La figure III.7 est un exemple d'extraction de la distribution des tensions de seuil effacées effectué sur un plan mémoire EEPROM.
Cette distribution, obtenue après un test de rétention (le véhicule de test, dans un état initial effacé, est maintenu à 250°C pendant 24 heures), met en évidence une queue de distribution qui correspond à une population de cellules mémoire qui semblerait perdre une partie des charges stockées dans la grille flottante. Ces premières informations électriques peuvent donner des indications relatives au type de défaillance affectant le plan mémoire. Cependant, les hypothèses avancées lors de cette étape de caractérisation électrique doivent être validées lors d'une analyse physique de la cellule (ou des cellules défaillantes).

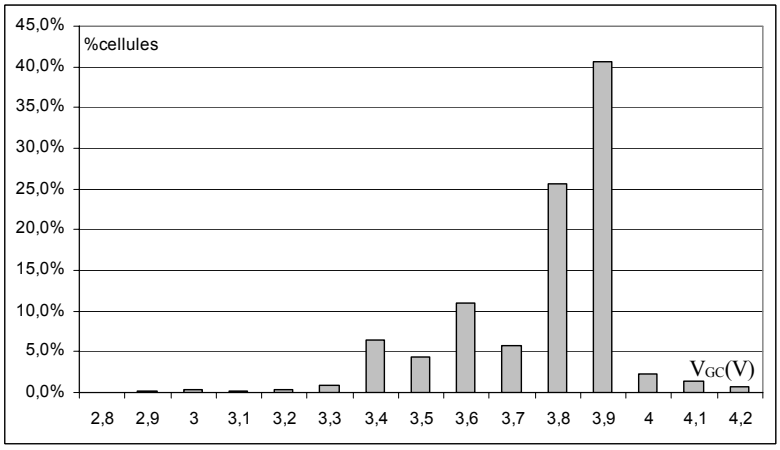


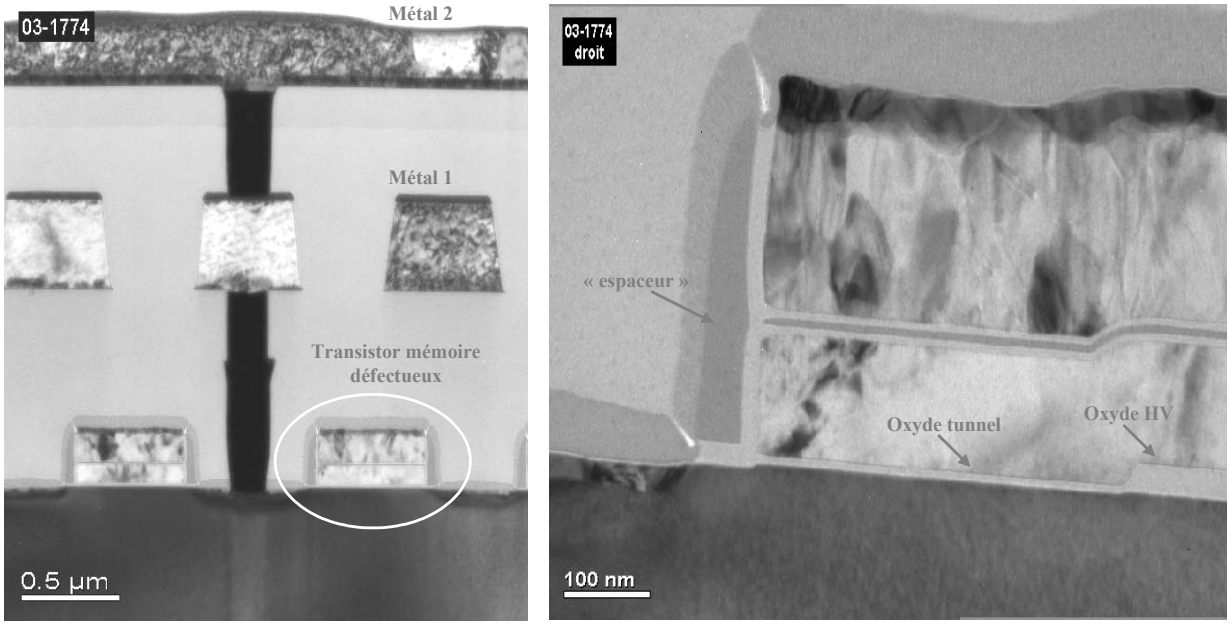
Figure III. 7 Distribution des tensions de seuil du plan mémoire EEPROM effacé.

d. Caractérisation physique

Dans le cadre de cet exemple, une analyse physique a été effectuée sur une cellule mémoire défaillante d'un véhicule de test EEPROM présentant des rejets en rétention.
La cellule défaillante, localisée à partir de la vue « bitmap » de la figure III.6a a été analysée de manière à tenter de déterminer l'origine physique de la défaillance. Lors d'une analyse physique, deux types d'opérations peuvent être réalisées :

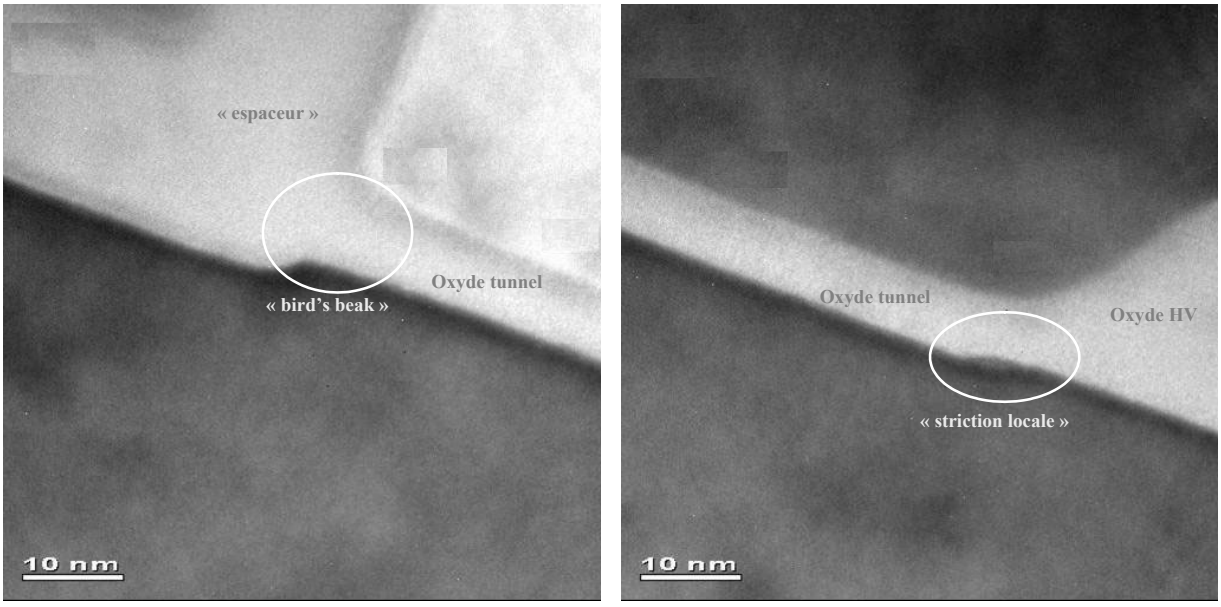
- la première consiste à effectuer des retraits de couches successifs jusqu'à remonter à une cause de défaillance (contacts défectueux ou niveaux de métallisation présentant des courts-circuits),
- la deuxième consiste à réaliser des coupes au niveau des cellules mémoires défaillantes.

La défaillance étant liée à un défaut de rétention, une coupe dans le sens du canal du transistor mémoire a été effectuée de manière à vérifier l'intégrité de l'oxyde tunnel.
Les figures III.8a et III.8b représentent une vue d'ensemble de la zone d'analyse : on distingue le transistor mémoire défectueux sur lequel une coupe transversale est réalisée. Cette coupe montre l'évolution de l'épaisseur de l'oxyde tunnel sur toute la longueur du canal du transistor mémoire. On peut noter que la préparation de l'échantillon à analyser a nécessité un retrait de couches jusqu'au niveau de métal 2.
Les figures III.9a et III.9b sont des vues rapprochées de la zone tunnel. On distingue les détails des deux extrémités de la fenêtre tunnel. Sur la figure III.9a, la transition entre l'« espaceur » et l'oxyde tunnel n'est pas franche (présence d'un bec d'oiseau ou « bird's beak »), ce qui entraîne une épaisseur d'oxyde tunnel anormalement élevée dans la zone de transition. La figure III.9b montre une striction locale au niveau de la zone de transition entre l'oxyde tunnel et l'oxyde HV. Ce pincement n'est pas présent sur les cellules témoins et pourrait causer un défaut de rétention.
De plus, des analyses effectuées sur les cellules voisines montrent une variation anormale de l'épaisseur de la couche d'oxyde tunnel (+/-10 Å) suivant le sens de la longueur. Cette dispersion des valeurs de l'épaisseur de l'oxyde tunnel pourrait aussi être à l'origine des rejets en rétention de plusieurs cellules du plan mémoire.



(a) (b)

Figure III. 8 Coupes transversales du point mémoire EEPROM.



(a) (b)

Figure III. 9 Vues rapprochées de la zone tunnel.

e. Conclusion

A ce stade, on s'aperçoit qu'une corrélation entre l'origine du défaut (étape du processus de fabrication mise en cause) et des signatures électriques pertinentes obtenues sur produit pourrait être à la base d'un outil de diagnostic de défauts dans le plan mémoire. Cet outil permettrait de remonter directement à la cause d'une défaillance sans passer par des études d'analyses de construction souvent coûteuses.
La suite de ce chapitre sera consacrée à l'introduction d'une méthodologie d'analyse de défaillance alternative permettant un diagnostic efficace des défauts dans le plan mémoire EEPROM.

II. Méthode d'analyse de défaillance alternative

A. Présentation et motivation

Les flots de tests fonctionnels des mémoires non volatiles de type EEPROM prennent en compte les temps de programmation longs inhérents à ce type de produit. En effet, les algorithmes utilisés sont des algorithmes simples de type n qui offrent une couverture de fautes limitée (cf. Chapitre I §III.A.3). De plus, à l'issue du test, les informations relatives à une puce défaillante ne permettent généralement pas de diagnostiquer l'origine de la défaillance.
Pour palier ce problème et éviter l'utilisation des techniques d'analyse de construction, une méthodologie de diagnostic de défauts basée sur des simulations électriques est présentée dans cette partie. Cette méthodologie devra permettre de déterminer la corrélation entre des signatures électriques de défauts simulés et l'étape du processus de fabrication responsable de l'apparition de ces défauts.
La mise en place de cette méthodologie de diagnostic passe par les étapes suivantes :

- la localisation et l'extraction des défauts potentiels pouvant affecter la mémoire. Cela passe par une étude du « Layout » et des niveaux de masques du circuit considéré,
- la modélisation électrique du comportement des défauts étudiés,
- l'insertion des modèles de défauts au sein du circuit mémoire,
- la simulation du comportement électrique du circuit mémoire incluant ces divers types de défauts,
- l'extraction des signatures électriques relatives aux défauts injectés.

En ce qui concerne les mémoires EEPROM, il apparaît clairement que la connaissance des tensions de seuil de toutes les cellules du plan mémoire est un paramètre clé dans le développement d'une méthodologie de diagnostic précise.
L'extraction d'une liste de fautes potentielles doit être effectuée pour une implémentation physique d'un circuit donné de manière à être efficace. Ainsi, la génération des fautes se base uniquement sur la représentation «layout » du circuit, ce qui permet d'obtenir une bibliothèque de fautes réaliste.

B. Architecture d'étude

1 Technologie étudiée

Dans le cadre de notre étude, nous considérerons la technologie EEPROM F6DP. Tout d'abord, un exposé sur le procédé de fabrication de cette technologie (schématisé par les figures III.10 et III.11) nous est nécessaire afin de mieux comprendre les mécanismes de défaillances associés à chaque étape du procédé de fabrication de ce type de mémoires EEPROM.

a. Procédé de fabrication « Front-End »

Le procédé de fabrication « Front-End » regroupe toutes les étapes liées à la fabrication des transistors :

- formation des caissons (figure III.10a),
- isolation des zones actives (figure III.10b),
- formation du canal, croissance des oxydes, dépôt du polysilicium de grille, ajustement des tensions de seuil… (figure III.10c et figure III.10d),
- définition des « espaceurs » et formation des zones de source et de drain des transistors (figure III.10e).

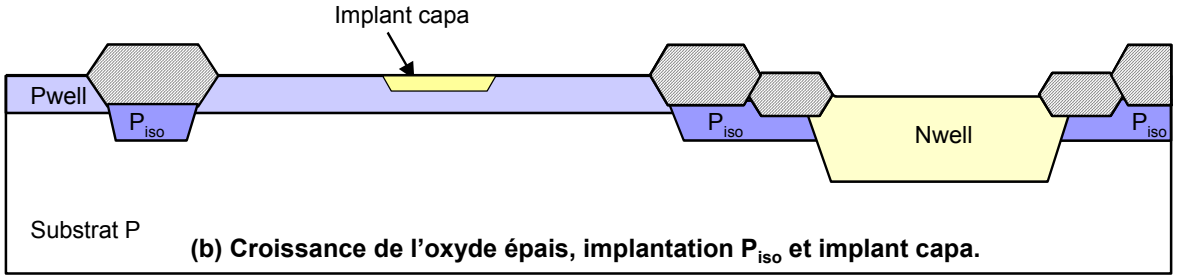
A partir d'un substrat de type P d'une résistivité de 2 Ω/cm, deux caissons de type N et P (nommés Nwell et Pwell) sur lesquels seront fabriqués les transistors de type P et N sont implantés à travers un oxyde initial (figure III.10a).
L'isolation des différents composants est effectuée par un procédé de type « Recessed LOCOS ». Ce procédé consiste à graver le substrat de silicium à une certaine profondeur, puis à faire croître un oxyde épais appelé oxyde de champ (ou « LOCOS ») d'une épaisseur de l'ordre de 600 nm. Cet oxyde enterré permettra d'isoler tous les composants afin d'éviter les fuites de courant entre les transistors adjacents. Le fait d'enterrer cet oxyde épais permet de réduire l'effet « bec d'oiseau » qui entraîne une diminution des géométries des transistors.
Une implantation de Bore au niveau des oxydes épais (implantation appelée P_{iso}) est ensuite réalisée de manière à renforcer l'isolation entre les zones actives des transistors NMOS. Une autre étape d'implantation de Phosphore à travers un oxyde sacrificiel, avec une dose de 5.10^{14} cm^{-2} permet la création des zones d' « implant capa » (zone d'injection tunnel et de certaines capacités périphériques) (figure III.10b).
Trois types d'oxydes sont présents dans le produit EEPROM final : l'oxyde HVOX (High Voltage Oxide), qui supporte les hautes tensions nécessaires à la programmation du plan mémoire (inférieures à 18V), l'oxyde LVOX (Low Voltage Oxide), qui supporte des tensions plus faibles (inférieures à 8V), et l'oxyde tunnel, à travers lequel se fait l'injection de charges dans la grille flottante. Pour fabriquer ces trois oxydes, on procède à trois oxydations successives :

- la première est une oxydation « sèche », avec un recuit à 900°C sous O2 et Dichloroéthylène (DCE). Le chlore présent dans le DCE permet de réduire les contaminations dans le four. L'épaisseur d'oxyde obtenue est de 16 nm. Cet oxyde est gravé dans les zones destinées à recevoir un oxyde final de type LVOX, ainsi que dans la zone tunnel;
- une deuxième oxydation a ensuite lieu et permet d'obtenir un oxyde de 18 nm. Cette oxydation est de type « humide » avec un recuit à 800°C sous H2, O2 et DCE. La présence d'hydrogène pendant le recuit permet d'obtenir un oxyde de meilleure qualité [OHM94];
- l'oxyde HV est enfin gravé au niveau de la zone d'injection, pour y faire croître l'oxyde tunnel. La recette de celui-ci est divisée en deux parties : une première oxydation humide a lieu à 770°C, puis une étape de nitruration est faite lors d'un recuit à 1000°C sous N_2O, pendant un temps très court. On obtient ainsi un oxyde tunnel dit « nitruré ». Il a été montré [CHE93] [DRG96] que les oxydes nitrurés offrent une meilleure tenue au claquage que les oxydes humides standard.

L'oxyde épais HVOX est obtenu par la combinaison des trois oxydations successives et son épaisseur finale est d'environ 35 nm. L'oxyde LVOX est le résultat de deux oxydations

successives (deuxième oxyde LVOX plus l'épaisseur de l'oxyde tunnel), et son épaisseur est d'environ 23 nm. Enfin, l'oxyde tunnel est présent uniquement dans la zone d'injection sur une zone fortement implantée (implant capa) et possède une épaisseur de 7 nm (figure III.10c). Il est également important de noter que les recuits qui ont lieu pendant les étapes d'oxydation servent également à la diffusion des dopants résultant des implantations précédentes (caissons N et P, P_{iso} et implant capa).
Cette étape de fabrication des oxydes est suivie d'une étape de définition des grilles des transistors.
Pour cela, une première couche de silicium amorphe (appelé Poly1) est déposée, avec une épaisseur de 140 nm. Elle constitue la grille flottante de la cellule mémoire. Une deuxième couche d'isolation inter polysilicium (couche ONO composée de trois matériaux : $S_iO_2/S_{i3}N_4/S_iO_2$) est ensuite déposée par un procédé CVD (Chemical Vapor Deposition). Cette couche d'isolation est gravée sélectivement pour permettre la connexion entre les grilles de polysilicium des transistors MOS périphériques. La grille de contrôle fortement dopée, est enfin définie par une couche de polysilicium (Poly2) d'épaisseur 140 nm plus une couche de Siliciure de Tungstène (WS_{i2}) d'épaisseur 250 nm. Cette dernière couche permet d'améliorer la résistivité de la grille.
L'empilement Poly1/ONO/Poly2 est ensuite gravé ce qui définit les dimensions des différents transistors (figure III.10d).
Une réoxydation a lieu pour densifier le siliciure et laisser une couche protectrice autour de l'empilement des différentes structures. Il s'agit d'un recuit à 900°C sous O_2. Des implantations à faible énergie sont faites à ce niveau du procédé, après masquage (implantation de BF_2 pour les transistors de type P, et de Phosphore pour les transistors de type N). On définit ensuite les « espaceurs » en déposant une couche d'oxyde, puis en la gravant de manière anisotrope.
Les implantations dites « Source/Drain » de type P^+ pour les transistors P-MOS, et N^+ pour les transistors N-MOS sont finalement réalisées et créent les zones de source et de drain des transistors (figure III.10e).

(a) Définition des caissons Nwell et Pwell.



(b) Croissance de l'oxyde épais, implantation P_{iso} et implant capa.

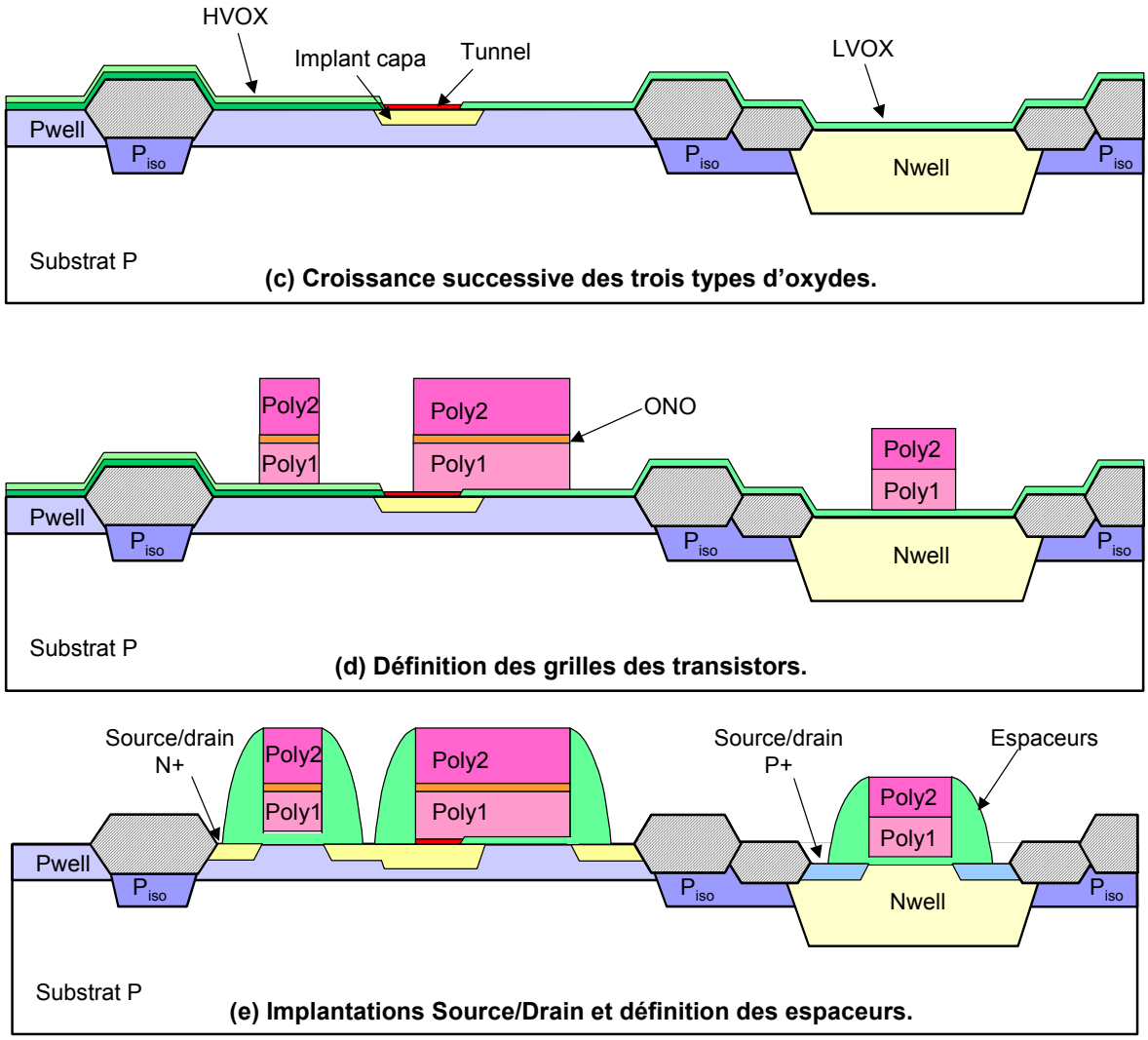


Figure III. 10 Formation des transistors (« Front-End »).

b. Procédé de fabrication « Back-End »

Le procédé de fabrication « Back-End » regroupe toutes les étapes de fabrication liées aux interconnexions :

- le dépôt des couches diélectriques de manière à isoler les composants et les couches de métal entre elles,
- la gravure des contacts et des vias,
- la passivation.

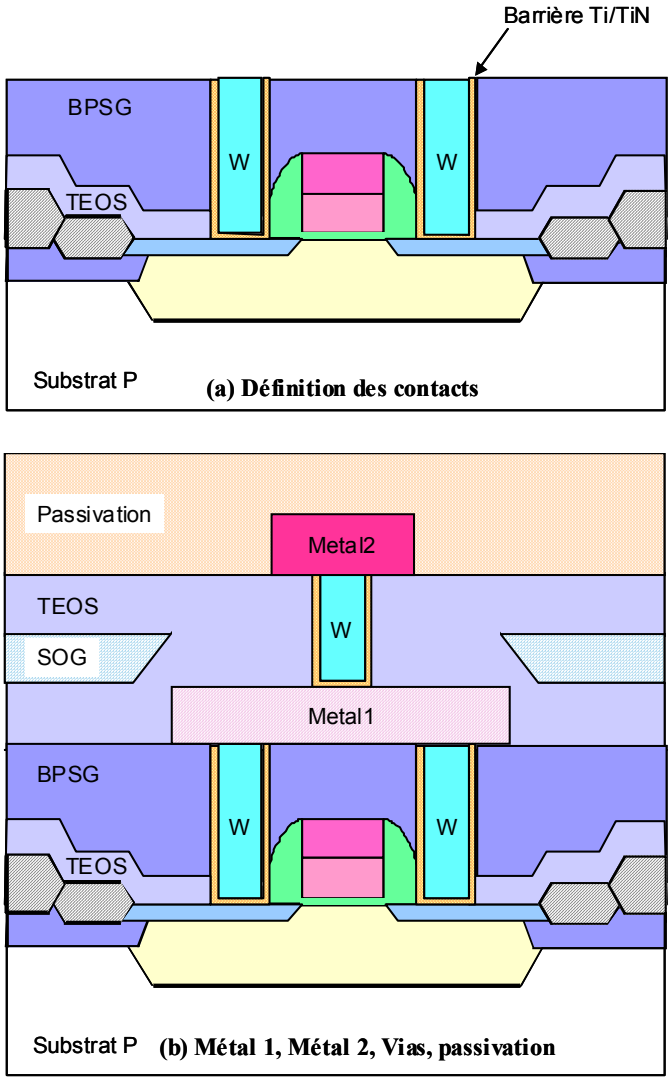


Figure III. 11 Formation des interconnexions (« Back-End »).

Ce procédé commence par le dépôt d'une couche de TEOS (TétraEthylOrtholoSilicate) de 100 nm, puis d'une couche de BPSG (Boron and Phosphorus doped Silicon Glass) d'environ 900 nm. Ensuite, une étape de recuit rapide RTA (Rapid Thermal Annealing) permet d'uniformiser la surface.
La définition des contacts consiste à graver les couches TEOS et BPSG. Une barrière Ti/TiN est ensuite déposée et le contact est rempli par du tungstène noté W (figure III.11a).
Une couche d'aluminium de 450 nm est ensuite déposée. Cette couche qui constitue le Métal 1 est gravée après masquage, jusqu'au niveau du BPSG.
Puis on dépose une épaisse couche de TEOS (550 nm), et on aplanit la surface par du SOG (Spin-On-Glass) qui sera gravé de manière anisotrope. Une deuxième couche de TEOS vient compléter ce que l'on nomme l'IMD (Inter-Metal Dielectric).
L'IMD est ensuite gravé pour définir une deuxième série de contacts appelés vias, dont le procédé de fabrication est identique à celui des premiers contacts.
Une deuxième couche d'Aluminium qui constitue le Métal 2 est déposée et gravée après masquage jusqu'au niveau de la couche TEOS.
Une couche de passivation servant à protéger la puce de l'environnement extérieur est finalement déposée. Cette couche finale est constituée d'USG (Undoped Silicon Glass) et de Nitrure. Notons que la couche de Nitrure est transparente aux rayons ultra-violets pour

permettre un effacement général des cellules mémoires. Le schéma final est représenté par la figure III.11b.

2 Architecture optimisée du plan mémoire EEPROM

a. Circuit de simulation

L'architecture générale du plan mémoire à simuler est représentée figure III.12. Les blocs fonctionnels qui forment l'architecture du plan mémoire EEPROM ont été décrits dans le chapitre I (§I.B.3). Une présentation générale des techniques de conception appliquée aux mémoires EEPROM est accessible à partir de la bibliographie [DAG03] [DAG00].
L'architecture du plan mémoire est de type NAND (cf. Chapitre I §I.B.2b). Ce plan mémoire, organisé en mots de 4 bits, est formé de 4 lignes et de 4 colonnes. Ce qui implique l'intégration dans le circuit d'une logique de décodage 4 bits (AD[0-3]). Les signaux haute tension sont transmis aux cellules mémoires par l'intermédiaire des bascules de lignes et de colonnes durant les opérations de programmation. Les données à programmer sont contenues dans un registre de données 4 bits. Les opérations de lecture sont effectuées en activant les amplificateurs de lecture qui fournissent la valeur logique du mot mémoire adressé (SA[0-3]).

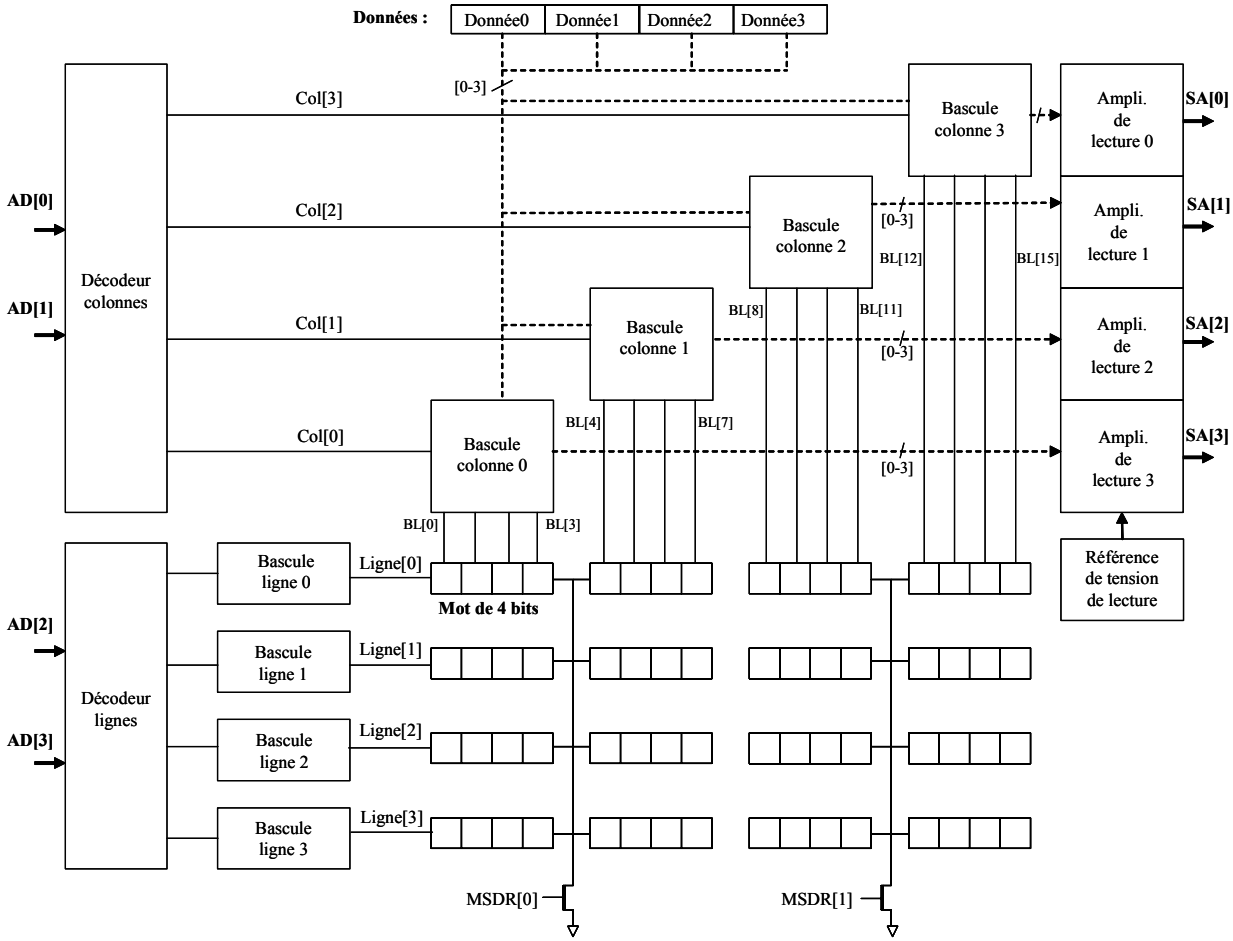


Figure III. 12 Architecture d'étude.

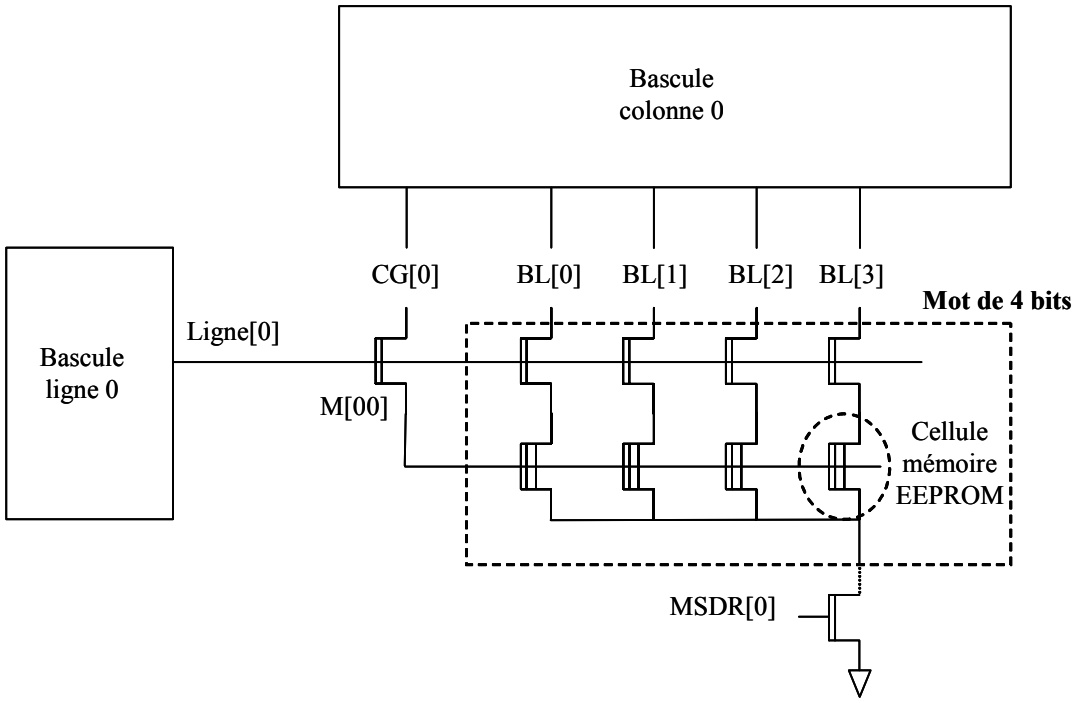


Figure III. 13 Mot mémoire formé de 4 bits.

La figure III.13 est une vue détaillée d'un mot mémoire (premier mot de la matrice mémoire, en haut à gauche du plan mémoire de la figure III.12). Ce mot mémoire est constitué de 4 transistors de sélection, commandés par le signal Ligne[0] et de 4 transistors à grille flottante. Les transistors de sélection permettent de transmettre la haute tension provenant des lignes de bits BL durant la phase d'écriture du mot mémoire. Le transistor de sélection de ligne, noté M[00], permet quant à lui de transmettre le signal haute tension CG[0] sur la grille des transistors à grille flottante durant l'effacement du mot mémoire quand la ligne Ligne[0] est active. Le signal MSDR[0] est actif durant la phase d'effacement du mot mémoire et à l'état logique bas durant la phase d'écriture du mot mémoire.

b. Temps de simulation

Un paramètre important à prendre en compte lors de la mise en place d'une méthodologie de diagnostic de défauts dans le plan mémoire EEPROM est le temps de simulation du circuit mémoire. En effet, les modèles de mémoires non volatiles se caractérisent par des temps de simulation longs. De ce fait, la taille de l'architecture d'étude doit refléter au maximum l'architecture du produit EEPROM tout en offrant des temps de simulation relativement courts. L'architecture étudiée est composée de 64 bits EEPROM modélisés en langage HDLA et offre des temps de simulations de l'ordre d'une heure.

3 Nature des défauts à simuler

Les défauts à insérer dans le circuit de simulation présenté figure III.12 sont spécifiques à la technologie utilisée. De manière à générer une liste de signatures électriques la plus exhaustive possible, les défauts introduits dans le circuit de simulation devront être de type résistif, de type capacitif, mais aussi modéliser l'influence des transistors parasites.

a. Notions d'aire critique

L'aire critique peut être définie comme une région du circuit physique (« layout ») dans laquelle la présence du centre d'une particule circulaire (contamination) entraînerait une défaillance. Les sources de ces défaillances sont classées en 3 catégories :

- matériau extérieur qui engendre des courts-circuits (figure III.14a),
- défaut de matériau qui engendre des circuits ouverts (figure III.14b),
- défauts au niveau des contacts (figure III.14c).

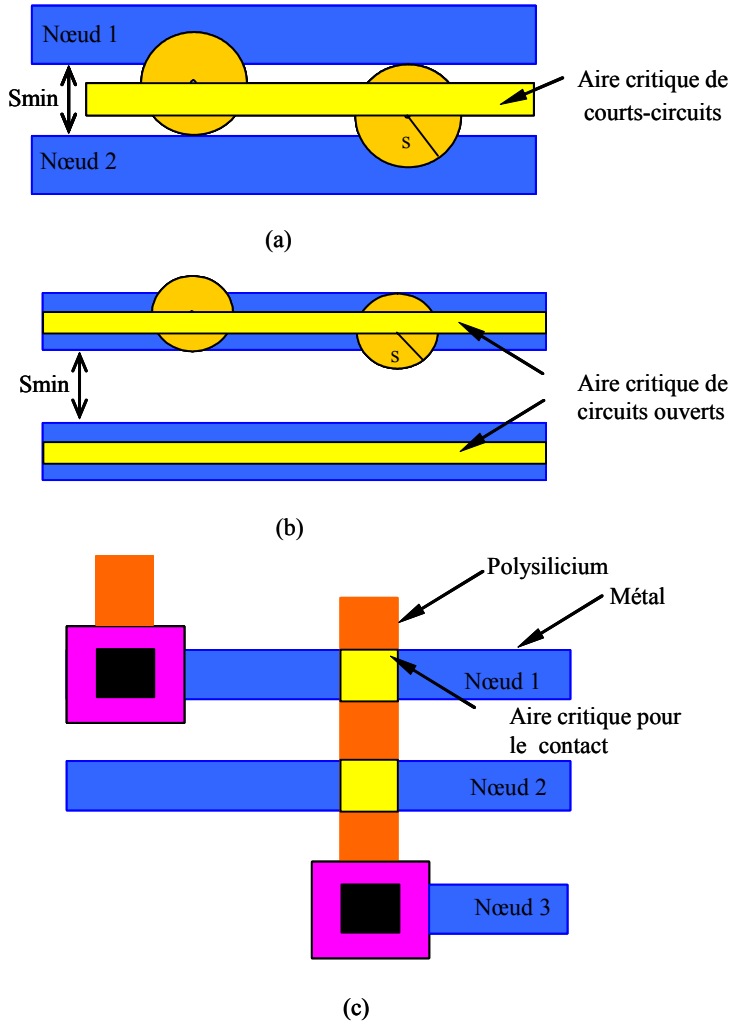


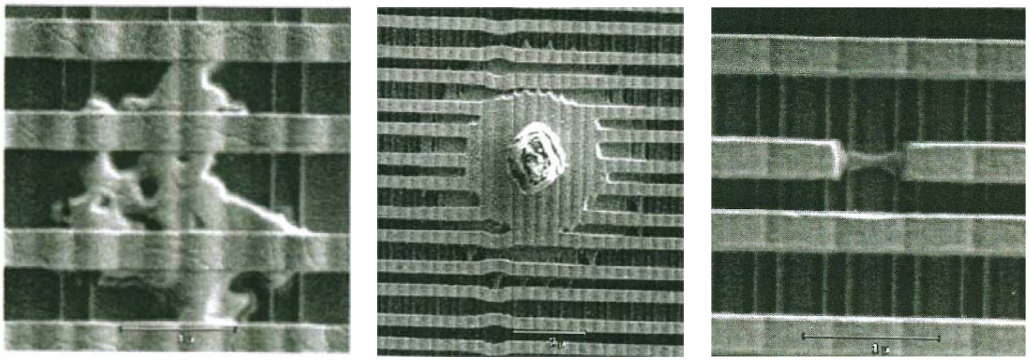
Figure III. 14 Aire critique.

KOREN propose une synthèse du concept d'aire critique et de ses applications [KOR01]. L'une des principales applications des extractions d'aire critique cible la modélisation de rendements [BAR98] [SEG00] [MIL99] [LEE97]. Des applications qui ciblent les circuits mémoires (SRAM et DRAM) ont été présentées par R. OTT [OTT99].
L'analyse de défaillance dans le plan mémoire EEPROM va passer par la prise en compte des notions d'aire critique, notamment pour l'introduction de défauts de type résistif dans le circuit de simulation.

b. Défauts de type résistif

Le plan mémoire utilise les règles de conception minimales pour une technologie donnée, ce qui le rend particulièrement sensible aux défauts de type particulaire.

Ainsi, de nouveaux problèmes liés principalement à la physique des interconnexions apparaissent (figure III.15). Les défauts particulaires peuvent être modélisés par des résistances variables (de 0.1KΩ à 100KΩ) de manière à simuler aussi bien l'influence d'un court-circuit entre lignes conductrices (figure III.15a et III.15b) que d'un circuit ouvert au niveau d'une ligne conductrice (figure III.15c).



(a) (b) (c)

Figure III. 15 contaminations (a)(b) et problème de motif (c) au niveau des lignes conductrices du plan mémoire.

c. Défauts de type capacitif

Par ailleurs, la diminution des distances entre les lignes conductrices crée des couplages diaphoniques entre des signaux logiquement indépendants. La diaphonie est le résultat d'un couplage capacitif et inductif entre deux lignes adjacentes. Une fraction de l'énergie de la ligne active peut être transférée sur la ligne non active par la capacitance et l'inductance mutuelles entre les deux lignes. Ainsi, les champs produits par les variations de tension et de courant sur une ligne peuvent induire tensions et courants sur la ligne voisine. De ce fait, des perturbations sont ainsi générées et le signal transmis dans l'une des deux lignes peut être détérioré ou influencé par la seconde.

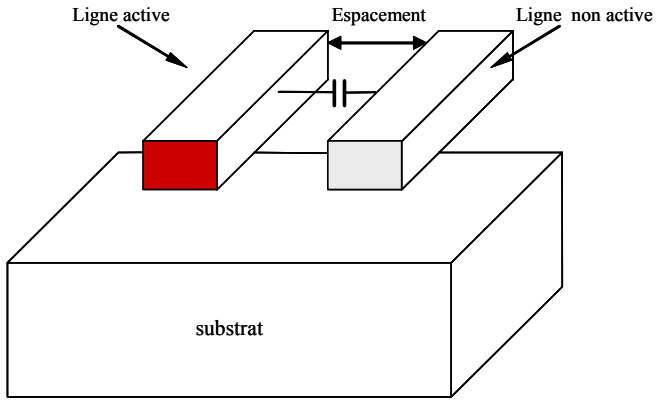


Figure III. 16 Couplage capacitif entre deux lignes conductrices.

Ce phénomène parasite se caractérise au niveau du plan mémoire EEPROM par un couplage capacitif entre deux lignes métalliques voisines (figure III.16). Ce phénomène est d'autant plus marqué que les mémoires EEPROM utilisent des signaux haute tension qui se caractérisent par des transitions brutales durant les phases de programmation.
La diaphonie dans les circuits intégrés a fait l'objet de nombreuses études. Le phénomène a tout d'abord été étudié pour les technologies CMOS [ROC94].

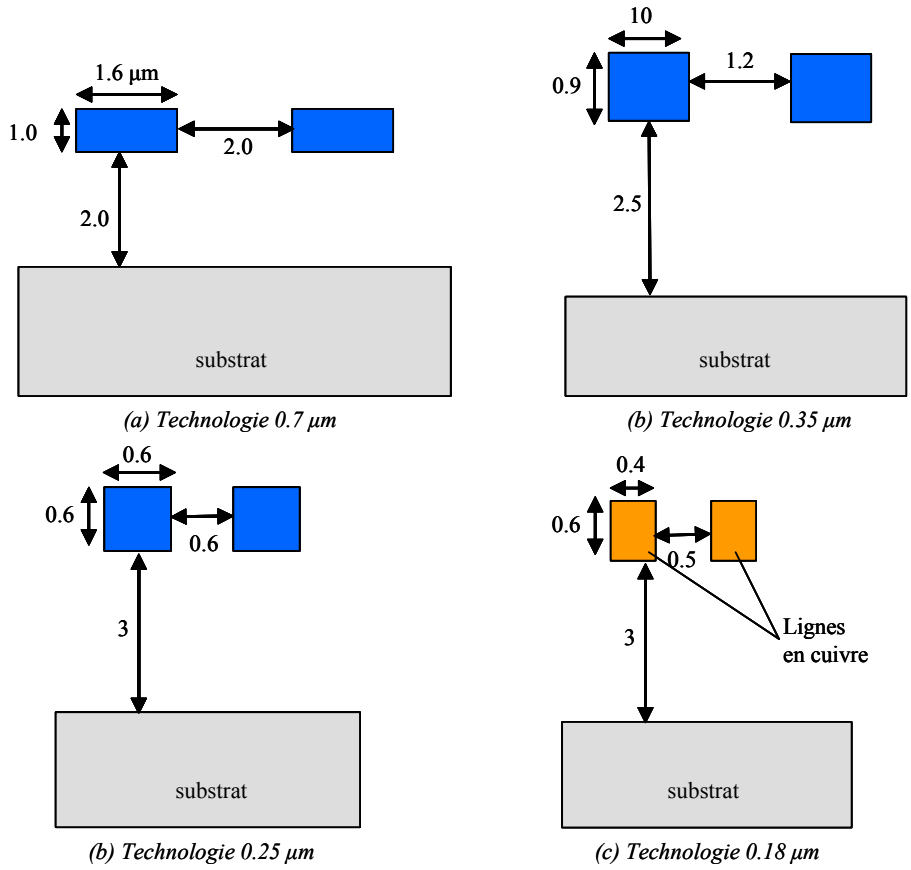


Figure III. 17 Evolution du couplage capacitif avec la technologie.

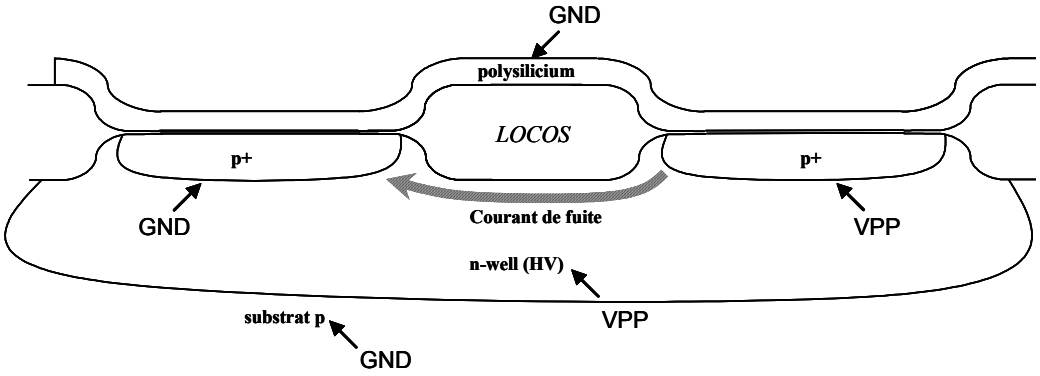
La figure III.17 donne une vue en coupe de lignes conductrices pour différentes technologies. Pour chaque technologie nous dessinons une largeur de métal typique du routage. Cette géométrie ne correspond pas aux règles minimales, en revanche elle est représentative des connexions routées automatiquement, qui constituent la quasi totalité des interconnexions [SIC98].

Pour les premières technologies (0,7 µm), on note une forme relativement plate des interconnexions métalliques, avec une large surface en regard avec le substrat et de faibles surfaces en regard des lignes entre elles, ce qui limite de ce fait les effets possibles de diaphonie par proximité entre pistes. L'interconnexion est donc principalement de nature capacitive, rapportée vers le substrat.

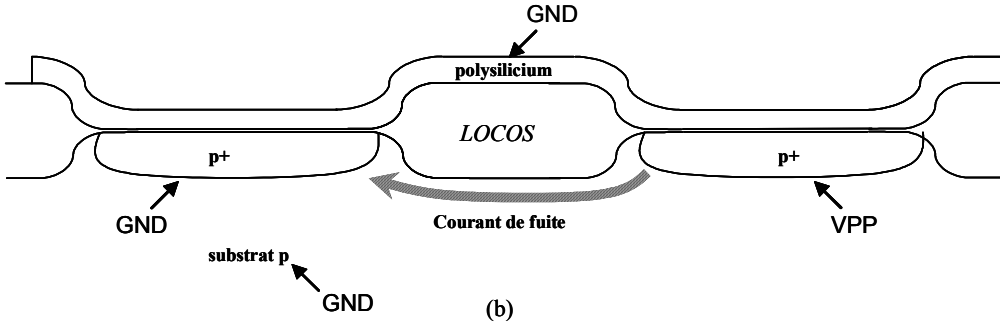
Pour les technologies plus avancées (0,25 µm et 0,18 µm), on note d'une part une faible surface en regard des lignes conductrices avec le substrat et d'autre part des surfaces en regard des lignes entre elles plus importantes. Dans ces conditions, le couplage diaphonique inter niveau de métallisation devient un problème majeur.

d. Transistors parasites

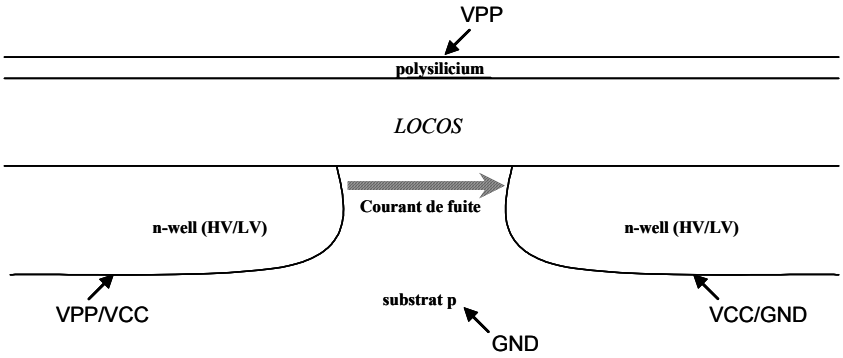
Une deuxième conséquence de l'utilisation des signaux haute tension est l'activation de composants actifs comme les transistors parasites. La figure III.18 répertorie les différents courants de fuite consécutifs à l'activation de transistors parasites. Ces courants de fuites (entre transistors théoriquement isolés) peuvent potentiellement perturber le fonctionnement des mémoires EEPROM. Pour chaque cas de figure, les potentiels à appliquer de manière à activer le transistor parasite considéré ainsi que le courant de fuite généré sont représentés.


GND
polysilicium
LOCOS
p+
p+
GND
Courant de fuite
VPP
n-well (HV)
substrat p
VPP
GND

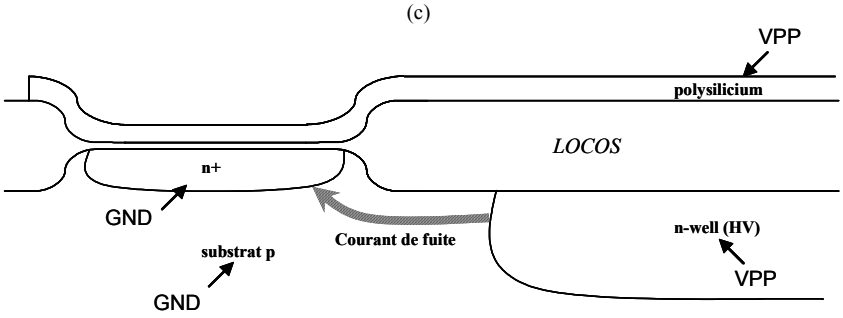
(a)


GND
polysilicium
LOCOS
p+
p+
GND
Courant de fuite
VPP
substrat p
GND

(b)


VPP
polysilicium
LOCOS
Courant de fuite
n-well (HV/LV)
n-well (HV/LV)
VPP/VCC
substrat p
GND
VCC/GND

(c)


VPP
polysilicium
LOCOS
n+
GND
Courant de fuite
n-well (HV)
substrat p
GND
VPP

(d)

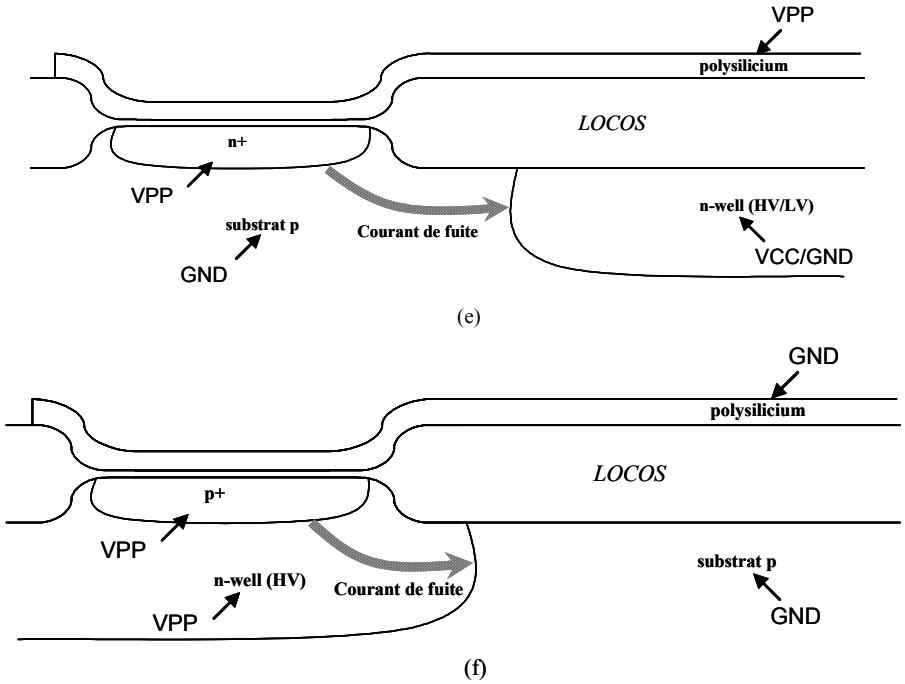


Figure III. 18 Courants de fuite générés par l'activation de transistors parasites.

C. Etudes des niveaux de masques (Technologie F6DP)

1 Niveaux de masques

Le nombre et la nature des masques utilisés définissent une technologie. L'étude des défauts relatifs à l'EEPROM se fera dans un premier temps en considérant chaque niveau de masque séparément. Une deuxième phase mettra en évidence les défauts affectant deux niveaux de masques différents.

Le tableau III.1 répertorie les différents masques de la technologie étudiée. Chaque masque est associé à un numéro qui correspond à l'étape de photolithographie dans laquelle est utilisé le masque. La technologie F6DP se base sur un procédé de fabrication CMOS classique. Les niveaux de masques relatifs à la partie mémoire EEPROM (masque implant capacitif numéroté 265, masque oxyde tunnel numéroté 405, masque de grille flottante numéroté 556, masque « Matrix » numéroté 525…) viennent compléter les masques relatifs à la technologie CMOS standard.

La polarité du masque indique la nature de la résine utilisée. En effet, durant l'étape de photolithographie, une fois la résine parfaitement étalée, la plaquette est insolée à l'aide d'une source lumineuse (lampe au Mercure) qui émet des ultraviolets. Dans le cas d'une résine positive, les parties qui sont insolées subissent une modification qui va rendre la résine beaucoup plus soluble que les autres. Dans le cas d'une résine négative, celle-ci est initialement soluble et c'est l'insolation qui durcit la résine non protégée. La colonne « Alignement » du tableau III.1 définit l'alignement du masque considéré rapport à un masque précédent.

Masque	Niveau Photo	Polarité	Alignement
Caissons N	015	**Négatif**	
Caissons P	055	Positif	015
Active	105	Positif	015
Isolation P	255	Positif	105
Implant Capacitif	265	Positif	105
Oxyde Grille HV	455	Positif	105
Oxyde Tunnel	405	**Négatif**	105
Grille Flottante	556	**Négatif**	105
LVS Array Implant	516	**Négatif**	105
Implant EPM	275	**Négatif**	105
Matrix	525	Positif	105
Native txtor II adj.	305	Positif	105
Polysillicium logique	505	Positif	105
Auto-alignement	515	**Négatif**	105
Auto-alignement Source	586	**Négatif**	505
Codage P-ROM	645	**Négatif**	105/505
Diffusion N+	605	Positif	105
Protection N+ et ROM	615	Positif	105
Diffusion P+	655	**Négatif**	105
Contact	705	**Négatif**	505
Contact N+ Plug II	735	Positif	105/705
Contact P+ Plug II	745	**Négatif**	105/705
Metal 1	800	Positif	705
Vias 1	850	**Négatif**	800
Metal 2	860	Clear	850
Vias 2	870	**Négatif**	860
Metal 3	880	Positif	870
Passivation	900	**Négatif**	705

Tableau III. 1 Niveaux de masques de la technologie F6DP.

2 Vue « Layout »

Une vue des dessins de masque d'une partie du plan mémoire EEPROM est présentée figure III.19. On distingue clairement les colonnes (BL[0-7]), les lignes (ROW[0-3]) ainsi que les lignes de source de cette partie de plan mémoire. L'extraction de fautes relatives à cette technologie se fera uniquement à partir du dessin de masques (et non au niveau portes logiques) de manière à obtenir une bibliothèque de fautes réaliste. Les masques seront affichés et analysés successivement dans l'ordre de leur utilisation dans le processus de fabrication.

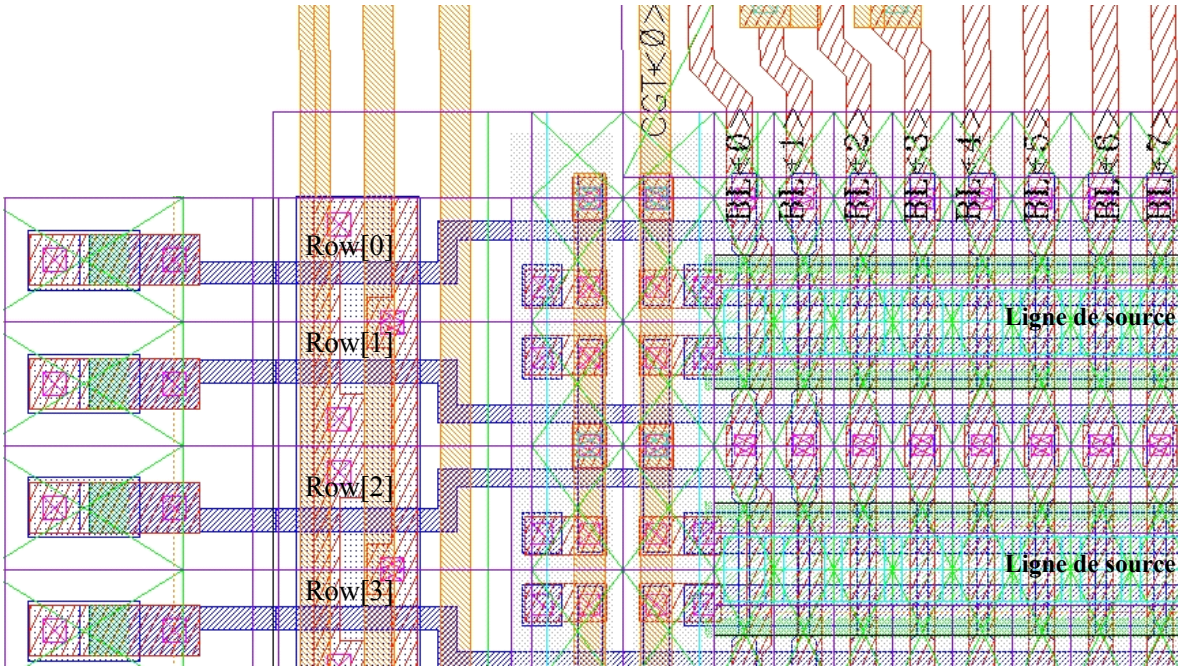


Figure III. 19 Vue « Layout » d'une partie de plan mémoire EEPROM.

Bien entendu, l'extraction des fautes pouvant potentiellement affecter le fonctionnement du circuit se fera à partir des modèles de défauts présentés précédemment en se basant sur les notions d'aire critique.

3 Défauts affectant les niveaux de masques

Les principaux défauts pouvant potentiellement affecter le fonctionnement du plan mémoire EEPROM seront présentés dans cette partie. Nous verrons que ces défauts peuvent engendrer des interactions entre deux cellules adjacentes ou affecter le fonctionnement d'une cellule mémoire par le biais d'interactions entre le transistor mémoire et le transistor de sélection. Ces défauts peuvent aussi affecter de manière indépendante le transistor de sélection ou le transistor mémoire d'une même cellule EEPROM.

a. « Layout » simplifié d'une cellule EEPROM unitaire

La figure III.20 est une représentation des principaux niveaux de masques d'une cellule EEPROM unitaire, formée du transistor de sélection et du transistor mémoire. Au niveau du drain du transistor mémoire (transistor d'état), on peut noter que la zone d'injection tunnel est définie par l'intersection des masques suivants : masque implant capacitif (masque 265), masque de grille flottante (masque 556) et masque définissant l'oxyde tunnel (masque 405). Le plan mémoire EEPROM est une répétition du motif élémentaire de la figure III.20.

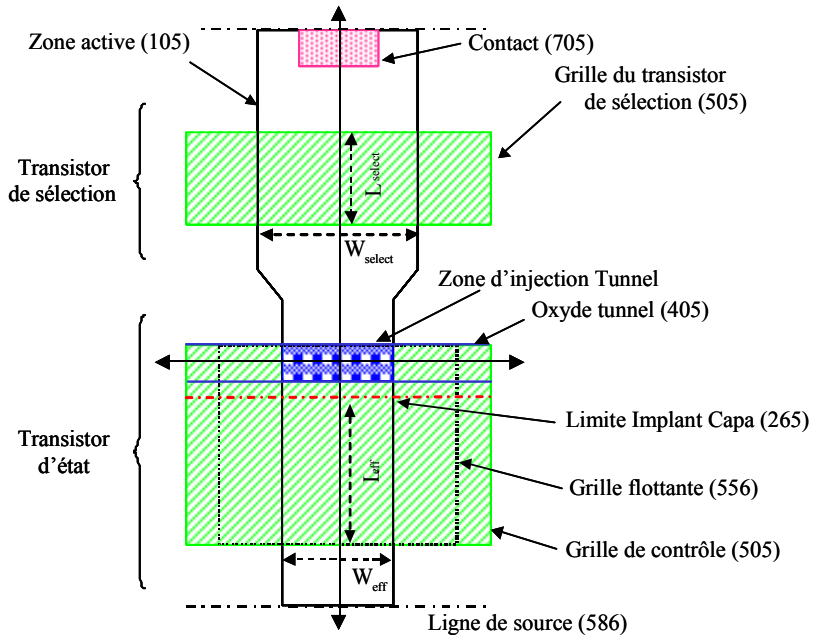


Figure III. 20 « Layout » simplifié d'une cellule mémoire unitaire.

b. Interactions entre cellules adjacentes

L'isolation des cellules mémoires élémentaires est réalisée par un procédé de fabrication compatible avec la technologie CMOS appelé LOCOS (cf. §II.B.1a), ce qui permet d'éviter les fuites de courant entre les transistors de zones actives adjacentes (figure III.21b). Ce procédé d'isolation entraîne une diffusion latérale de l'oxyde de champ au détriment de la surface occupée par la zone active. Ce phénomène parasite a fait l'objet de nombreuses recherches [APP70] [WU83] [FAL91] et se traduit par une perte de surface silicium (réduction de la zone active) qui entraîne une réduction des largeurs W_{select} et W_{eff} et une détérioration des caractéristiques électriques des transistors.
De manière à renforcer l'isolation entre les zones actives, une implantation de bore est réalisée sous l'oxyde épais. Les défaillances typiques relatives à cette étape sont l'apparition de courants de fuite entre lignes de bit provoquées par l'apparition de transistors parasites, aussi bien au niveau des transistors de sélection, qu'au niveau des transistors mémoire. De plus, une implantation insuffisante de bore peut entraîner l'apparition de résistances parasites entre les lignes de bit.
La ligne de source est commune aux bits de deux mots mémoires adjacents. La ligne de source, représentée figure III.19, est dans un premier temps gravée (masque 586). Puis une étape d'implantation va rendre cette ligne la plus conductrice possible. Cependant, il existe une résistance de ligne de source qui perturbe les phases de lecture du mot mémoire. Cette résistance conduit à une diminution du courant de drain et à un décalage positif des tensions de seuil. Elle a été mesurée à 2.3 kΩ pour un mot de 8 bits (soit 290 Ω par bit) [ECS02].
La densité d'intégration élevée dans le plan mémoire peut conduire à des problèmes de courts-circuits entre les lignes de métaux (figure III.19) qui définissent les lignes (ROW) ou les colonnes (COL) du plan mémoire. De plus, de nombreuses analyses ont montré que des courts-circuits entre grilles flottantes (filament de polysilicium après gravure) pouvaient aussi affecter les phases de programmation du plan mémoire.
Des phénomènes de programmation parasites peuvent aussi apparaître entre deux cellules adjacentes par couplage capacitif entre deux lignes conductrices (activation d'une ligne ou d'une colonne non sélectionnée).

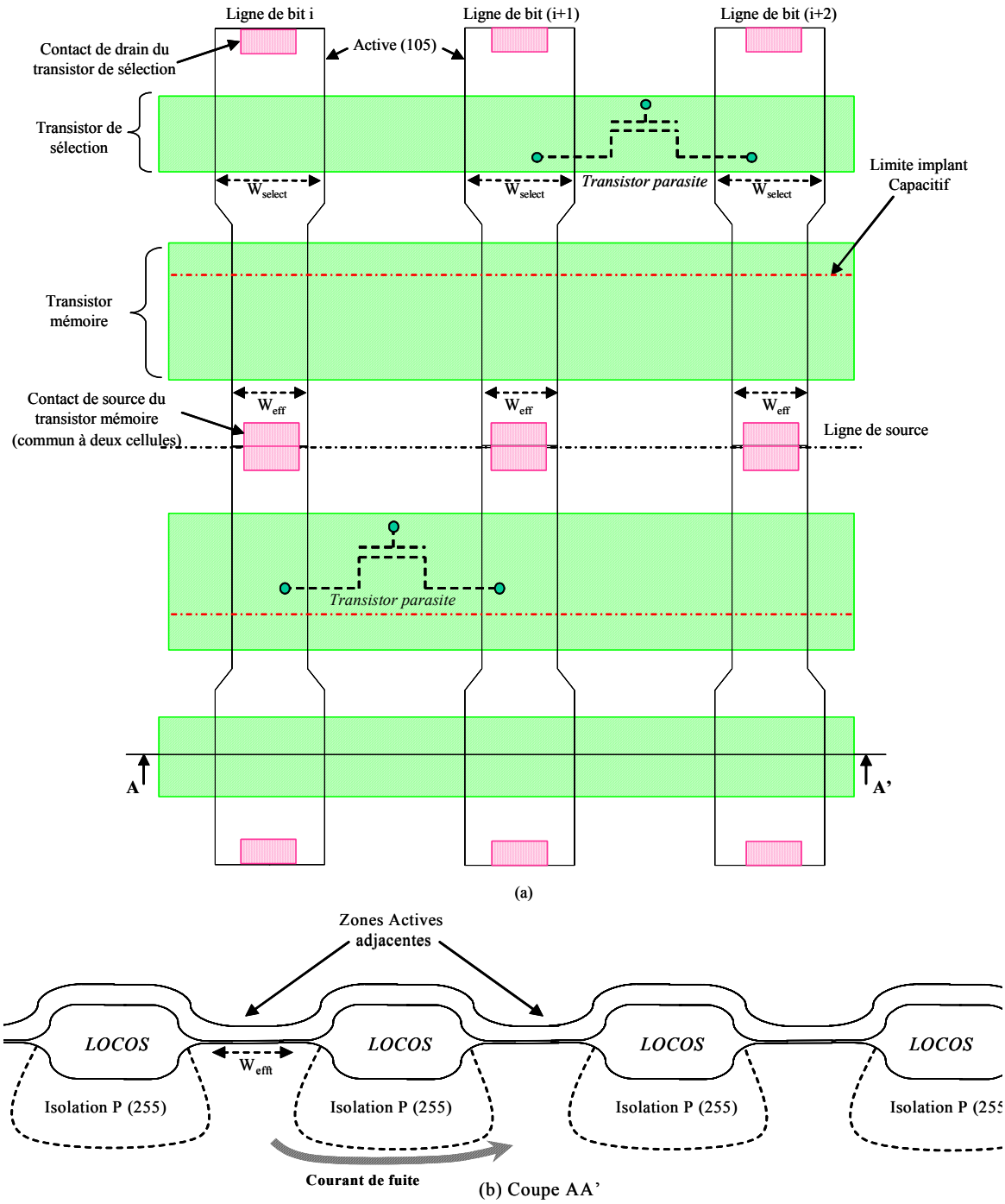


Figure III. 21 Interactions entre cellules mémoires adjacentes.

Au niveau des contacts (masque 700), la principale cause de défaillance se situe au niveau des contacts de drain du transistor de sélection. Ces contacts peuvent être très résistifs, ce qui induit une dégradation du signal haute tension durant les phases d'écriture ou trop profond (traverse le substrat), ce qui provoque dans ce cas une défaillance au niveau de toute une colonne du plan mémoire.

Un dernier type de défaut, plus rarement rencontré, touche les contacts de source et peut provoquer une défaillance au niveau de deux cellules adjacentes.

c. Interactions entre transistors d'un même bit

La zone d'injection est la fenêtre par laquelle les charges sont injectées dans la grille flottante. Elle se situe au niveau du drain du transistor mémoire. Ce dernier correspond à une implantation de type N^+ appelée « implant capacitif » (figure III.22). Le masque 265 va définir la taille de la zone dopée située au dessous de l'oxyde tunnel. On peut également noter que les zones implantées du drain du transistor d'état et de la source du transistor de sélection sont communes.
Cet implant capacitif définit la longueur de canal effective L_{eff} du transistor mémoire. De même, l'implant capacitif au niveau de la source du transistor de sélection définit la longueur de canal effective L_{select} du transistor de sélection. Ainsi, une implantation trop forte ou trop faible aurait un impact direct sur le dimensionnement des longueurs de la cellule mémoire. De plus, une implantation trop faible induirait une résistance entre la source du transistor de sélection et le transistor mémoire qui aurait pour conséquence une dégradation du signal de programmation durant les phases d'écriture.

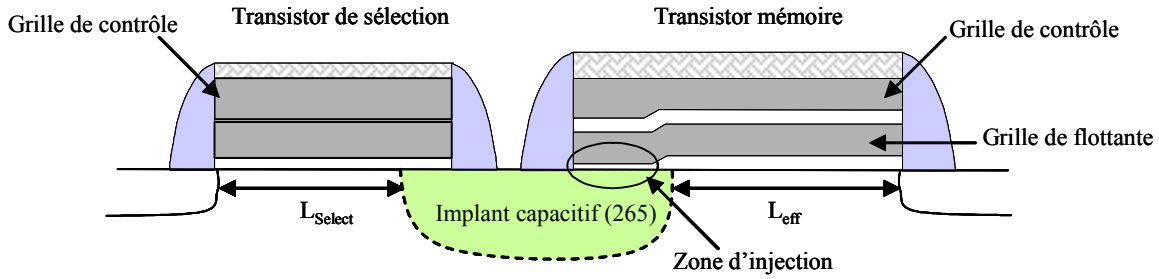


Figure III. 22 « Implant capacitif ».

La gravure des grilles de contrôle (masque 505) peut entraîner des courts-circuits entre le transistor mémoire et le transistor de sélection d'une même cellule. D'autres cas de courts-circuits entre la grille de contrôle et la grille flottante ont été relevés.
Les défaillances relatives à l'oxyde haute tension HV (masque 455) et l'oxyde tunnel (masque 405) sont nombreuses. En effet, comme nous l'avons vu, l'oxyde tunnel est obtenu après gravure de l'oxyde HV, puis déposition d'un oxyde LV et gravure de ce dernier au niveau de la zone d'injection tunnel. Les principales défaillances sont dues à un déplacement de la fenêtre tunnel suite à un désalignement entre le masque tunnel (masque 405) et le masque Active (masque 105). Ce désalignement induit des variations au niveau de la longueur de la fenêtre tunnel L_{tun} et par conséquent de la surface de la fenêtre tunnel S_{tun}.

d. Défauts affectant un transistor isolé

Les défaillances pouvant affecter les transistors isolés sont nombreuses. Elles ont généralement pour conséquence une variation des paramètres géométriques de la cellule mémoire (transistor mémoire et transistor de sélection). L'impact de la variation d'un paramètre géométrique du transistor mémoire sur les tensions de seuil a été traité au Chapitre II. Concernant le transistor de sélection de la cellule mémoire, l'utilisation de tensions de programmation élevées (~13V) peut entraîner le claquage de la jonction drain-substrat en écriture. En effacement, on peut observer un phénomène de claquage de l'oxyde de grille (GOS : Gate Oxide Short) de ce même transistor. Ces deux types de défaillances entraînent des courts-circuits ligne-colonne. Cette constatation est aussi valable pour le transistor de sélection d'un mot mémoire.

e. Table des interactions

Il serait très difficile d'établir une liste exhaustive de toutes les défaillances pouvant affecter le fonctionnement du plan mémoire EEPROM. En effet, la complexité du processus de fabrication ne permet pas de prendre en compte tous les scénarii de défaillances, même pour un seul niveau de masque. De plus, des interactions entre niveaux de masques différents peuvent apparaître. La table des interactions du tableau III.2 montre toute la complexité du problème. On peut noter que les interactions entre les différents niveaux de masques sont pondérées (aucune interactions, interactions fortes….), ce qui permet d'ignorer certaines interactions, jugées improbables. Cette table a été réalisée en collaboration avec les personnes en charge de l'introduction de la technologie F6DP.

	15/ 55	105	255	265	455	405	275	515	556	525	305	505	586	645	605	615	655	705	735	745	800	850	860	900
15/ 55	x	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na
105		x	ooo	o	o	o	na	x	o	x	x	o	o	na	na	na	na	o	o	na	na	na	na	na
255			x	x	o	o	x	x	x	x	x	o	x	x	x	x	x	x	x	x	x	x	x	x
265				x	o	o	x	x	o	x	x	oo	o	x	x	x	x	x	x	x	x	x	x	x
455					x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
405						x	x	ooo	x	x	x	ooo	x	x	x	x	x	x	x	x	x	x	x	x
275							x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
515								x	o	oo	x	x	x	x	x	x	x	x	x	x	x	x	x	x
556									x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
525										x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
305											x	x	x	x	x	x	x	x	x	x	x	x	x	x
505												x	o	x	x	x	x	x	x	x	x	x	x	x
586													x	o	oo	?	?	x	x	x	x	x	x	x
645														x	o	?	?	x	x	x	x	x	x	x
605															x	o	o	o	x	x	x	x	x	x
615																x	o	?	x	x	x	x	x	x
655																	x	oo	x	oo	x	x	x	x
705																		x	o	o	o	o	o	x
735																			x	o	o	o	o	x
745																				x	o	o	o	x
800																					x	o	o	x
850																						x	o	x
860																							x	x
900																								x

x : aucune interaction, na : aucune signification, 0 : faibles interactions, 00 : interactions modérées, 000 : interactions fortes, ? : A déterminer

Tableau III. 2 Table des interactions.

D. Injection de défauts dans le circuit de simulation

1 Algorithmes de test

Le flot de tests fonctionnels des mémoires EEPROM consiste soit à placer une même donnée dans toute la mémoire, soit à écrire des données complémentaires dans des cellules topologiquement adjacentes (« checkerboard »). Ces algorithmes de test seront réutilisés pour notre étude. L'algorithme qui consiste à écrire une diagonale dans le plan mémoire ne sera pas pris en compte puisque ce dernier cible la circuiterie de décodage et non le plan mémoire.
Ainsi, pour chaque type de défaut injecté dans le circuit de simulation, les algorithmes de test suivants seront appliqués :

- effacement du plan mémoire, puis lecture des cellules effacées (test All_00),
- écriture du plan mémoire, puis lecture des cellules écrites (test All_11),
- écriture du motif « checkerboard » direct, puis lecture de ce motif (CHK),
- écriture du motif « checkerboard » inverse, puis lecture de ce motif (CHKI).

2 Réponses du circuit aux défauts injectés

a. Valeurs logiques de sorties

Dans un premier temps, l'impact de chaque défaut simulé sera évalué durant la phase de lecture du plan mémoire en comparant les données de lecture attendues à l'état logique du mot mémoire adressé. L'état logique du mot mémoire est accessible sur les sorties SA[0-3] des amplificateurs de lecture (figure III.12).

b. Valeurs des tensions de seuil de chaque cellule du plan mémoire

Un deuxième paramètre à prendre en compte lors de la phase de lecture est la tension de seuil de chaque cellule EEPROM, et cela, pour chaque état électrique du plan mémoire. Ainsi, la connaissance des trois tensions de seuil d'une cellule défaillante pourra nous apporter des informations supplémentaires de manière à trouver une corrélation entre le défaut simulé et une variation anormale des tensions de seuil.

c. Outil logiciel

Les temps de simulation longs ont nécessité le développement d'un programme d'automatisation des simulations. Ce programme, que nous avons développé en langage de programmation PERL, nécessite en entrée un circuit de simulation décrit sous forme de « netlist » et un fichier de défauts qui regroupe tous les modèles défauts à simuler (figure III.23). Le programme principal insère successivement chaque défaut dans le circuit et extrait les réponses souhaitées après simulation (valeurs logiques et tensions de seuil).
Le fichier de sortie « résultp.txt », qui se présente sous forme d'un fichier texte, débute par la description du défaut inséré. Dans l'exemple de la figure III.24, il s'agit d'une résistance notée Rd_BL01_01 de valeur 0,1 kΩ insérée entre les nœuds électriques BL[0] et BL[1] du circuit de simulation.
A l'issue de la simulation du circuit incluant le défaut, les trois tensions de seuil sont relevées (ETAT VIRGE, EFFACE et ECRIT), ainsi que les états logiques des mots adressés (cinq mots dans cet exemple : MOT0 à MOT4).

Pour chaque ligne du fichier de sortie, la valeur numérique suivie du mot clé X_VALUE représente le temps ou l'extraction des réponses a été effectuée durant les simulations transitoires.

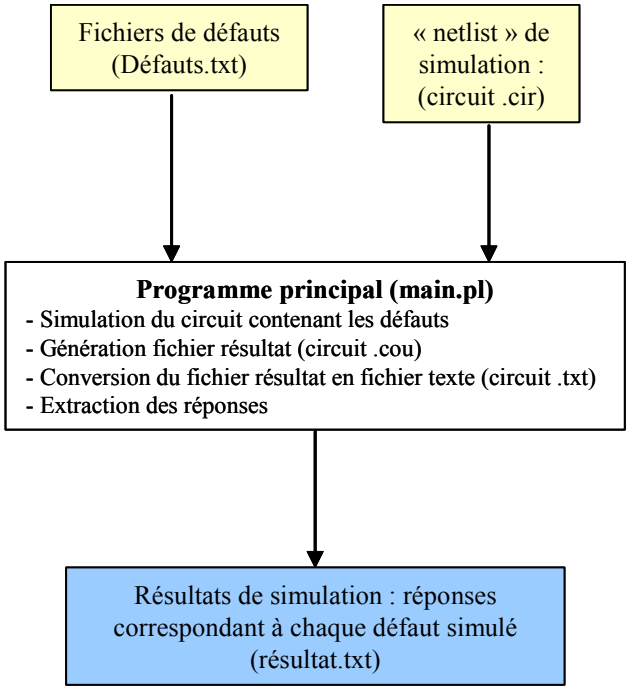


Figure III. 23 Outil logiciel.

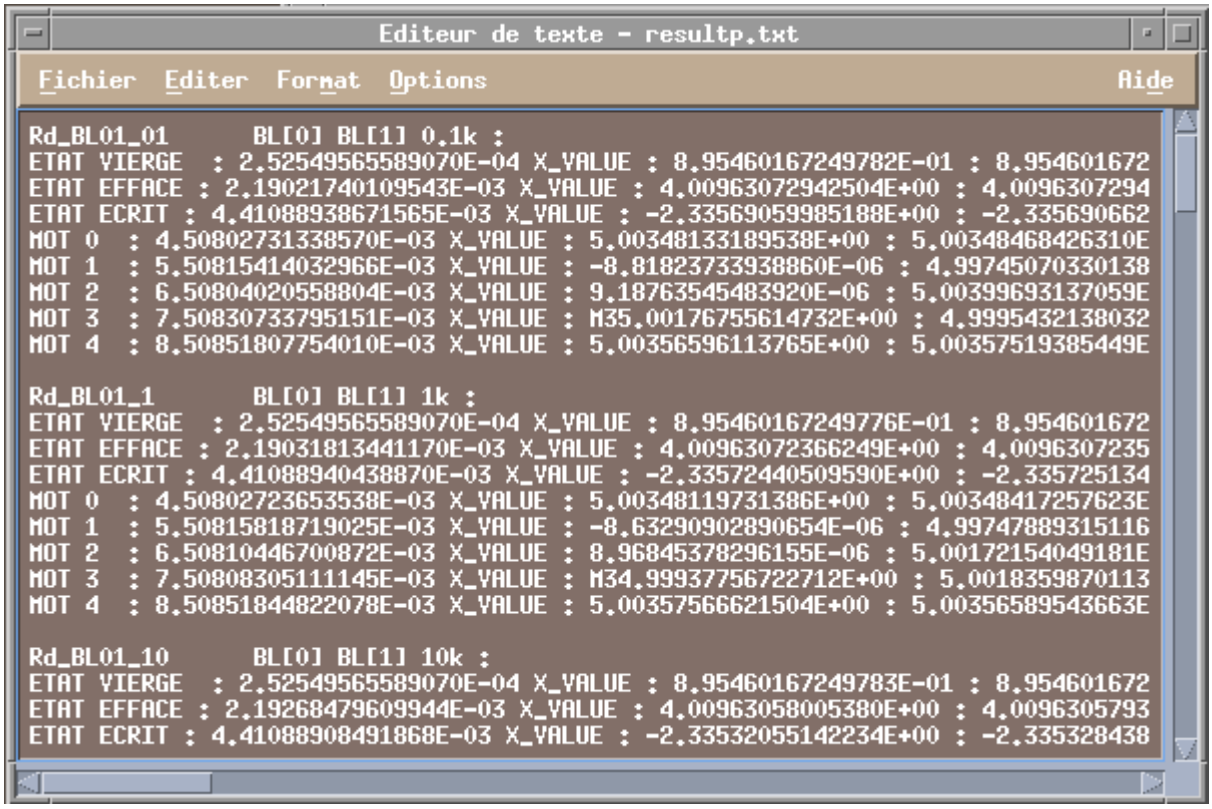


Figure III. 24 Fenêtre de sortie.

III. Résultats de simulation

A. Impact des défauts résistifs sur le plan mémoire

Le tableau III.3 est une représentation du plan mémoire EEPROM où chaque case est associée à une cellule élémentaire. Cette figure schématise le motif de test CHK à inscrire dans le plan mémoire. Le tableau III.4 représente les tensions de seuil finales extraites suite à l'inscription du motif CHK pour un circuit exempt de tout défaut.
Dans l'exemple que nous allons développer, le défaut injecté est une résistance de 0,1 kΩ située entre les nœuds BL[3] et BL[4] du circuit de simulation de la figure III.12. Le but étant de déterminer l'impact d'un court-circuit franc entre 2 lignes de bit sur les réponses du circuit.
Dans le cadre de cet exemple,

- le test All_00 ne montre aucune défaillance au niveau logique, les tensions de seuil effacées finales sont égales à 4,010 V (valeurs nominales),
- le test All_11 ne montre aucune défaillance au niveau logique, les tensions de seuil écrites finales sont égales à -2,357 V (valeurs nominales),
- le test CHK engendre une défaillance; la signature électrique correspondante touche deux colonnes défaillantes (tableau III.5) et les valeurs des tensions de seuil finales sont données par le tableau III.6,
- le test CHKI engendre lui aussi une défaillance, la signature électrique correspondante touche là aussi deux colonnes défaillantes et les valeurs des tensions de seuil finales présentent globalement les mêmes variations que le test CHK.

0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

Tableau III. 3 Motif de test CHK.

4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	4.010	-2.357	-2.357
-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	4.010	-2.357
4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	4.010	-2.357	-2.357
-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	4.010	-2.357

Tableau III. 4 Tensions de seuil représentatives du motif CHK.

0	1	0	1	1	1	0	1	0	1	0	1	0	1	0	1
1	0	1	1	1	0	1	0	1	0	1	0	1	0	1	0
0	1	0	1	1	1	0	1	0	1	0	1	0	1	0	1
1	0	1	1	1	0	1	0	1	0	1	0	1	0	1	0

Tableau III. 5 Résultats logiques après le test CHK.

4.010	-2.357	4.010	-2.340	-2.340	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357
-2.357	4.010	-2.357	-2.340	-2.340	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010
4.010	-2.357	4.010	-2.340	-2.340	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357
-2.357	4.010	-2.357	-2.340	-2.340	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010

Tableau III. 6 Extraction des tensions de seuil après le test CHK.

La représentation topologique des valeurs des tensions de seuil du plan mémoire EEPROM montre que les cellules défaillantes lues se trouvent dans un état écrit (V_T = -2,340 V) alors que l'état électrique attendu est un état effacé (V_T = 4,009 V).
On peut noter que les tensions de seuil des colonnes 3 et 4 du plan mémoire sont identiques à celles des cellules défaillantes, ce qui signifie que le court-circuit a pour conséquence un couplage entre les lignes de bit BL[3] et BL[4].

B. Impact des défauts capacitifs sur le plan mémoire

Dans cet exemple, le défaut injecté est une capacité de 100 fF située entre les nœuds BL[3] et BL[4] du circuit de simulation de la figure III.12. Le but visé est de mettre en évidence l'effet d'un couplage capacitif, affectant deux lignes de bit, sur les réponses du circuit. Cette fois, le motif de test CHKI (illustré tableau III.7) est inscrit dans le plan mémoire. Les tensions de seuil qui correspondent à ce motif sont données par le tableau III.8. Là aussi,

- le test All_00 ne montre aucune défaillance au niveau logique, les tensions de seuil effacées finales sont égales à 4,009 V,
- le test All_11 ne montre aucune défaillance au niveau logique, les tensions de seuil écrites finales sont égales à -2,352 V,
- le test CHKI engendre une défaillance, la signature électrique correspondante touche deux colonnes défaillantes (tableau III.9) et les valeurs des tensions de seuil finales sont données par le tableau III.10,
- le test CHK engendre lui aussi une défaillance et les tensions de seuil finales présentent globalement les mêmes variations que le test CHKI.

1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

Tableau III. 7 Motif de test CHKI.

-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010
4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357
-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010
4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357

Tableau III. 8 Tensions de seuil représentatives du motif CHKI.

1	0	1	1	1	0	1	0	1	0	1	0	1	0	1	0
0	1	0	1	1	1	0	1	0	1	0	1	0	1	0	1
1	0	1	1	1	0	1	0	1	0	1	0	1	0	1	0
0	1	0	1	1	1	0	1	0	1	0	1	0	1	0	1

Tableau III. 9 Extractions des tensions de seuil effacées.

-2.357	4.010	-2.357	-1.009	-2.345	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010
4.010	-2.357	4.010	-2.355	-1.017	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357
-2.357	4.010	-2.357	-1.009	-2.345	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.365	4.010	-2.357	4.010
4.010	-2.357	4.010	-2.355	-1.017	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357

Tableau III. 10 Extraction des tensions de seuil après le test CHKI.

C. Impact des transistors parasites sur le plan mémoire

Dans l'exemple qui suit, le défaut injecté est un transistor parasite situé entre les nœuds BL[0] et BL[1] du circuit de simulation de la figure III.12. Ce transistor parasite est activé par le potentiel appliqué sur la première ligne Ligne[0] de la matrice mémoire durant la phase d'effacement. Les résultats relatifs au test CHKI sont donnés par les tableaux III.11 et III.12.

1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

Tableau III. 11 Résultats logiques après le test CHKI.

-2.127	-2.341	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	4.010
-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010
-2.127	-2.341	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	4.010
-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010	-2.357	4.010

Tableau III. 12 Extraction des tensions de seuil après le test CHKI.

D. Analyses des résultats de simulation

1 Construction d'une base de données de signatures électriques

Les cellules défaillantes (présentant un état logique différent de celui attendu) ou marginales (présentant des tensions de seuil éloignées des valeurs nominales) sont répertoriées de manière à construire une base de donnée qui lie d'une part, les défauts simulés (résistances, capacités, transistors parasites) et d'autre part, les signatures électriques obtenues à l'issue du test (tensions de seuil, réponses logiques, parties du plan mémoire où la défaillance est localisée).
Les tests fonctionnels sont appliqués dans l'ordre suivant : test ALL_00, test All_11, test CHK, et test CHKI. Ainsi, pour ce flot de test, le premier des quatre tests à induire une défaillance est relevé et vient compléter les réponses relatives à chaque défaut injecté.

2 Exemple

Le tableau III.13 représente un extrait de la base de donnée obtenue après traitement de tous les types de défauts pouvant affecter le circuit mémoire.

Défaut	$V_{Tefface}$	V_{Tecrit}	$V_{Tvierge}$	Logique	Signature	Test
Rd_BL3_BL4_01kΩ	-2.340	- 2.340	0.895	FAIL	2 colonnes	CHK/CHKI
Rd_BL3_BL4_10 kΩ	-2.332	- 2.332	0.895	FAIL	2 colonnes	CHK/CHKI
Rd_BL3_BL4_10 kΩ	-2.321	- 2.321	0.895	FAIL	2lignes	CHK/CHKI
Capd_BL3_BL4_01fF	4.010	-2.322	0.896	PASS	-	CHK/CHKI
Capd_BL3_BL4_100fF	-1.009	-2.355	0.896	FAIL	2colonnes	CHK/CHKI
Tx_ BL1_BL2_Row	-2.127	-2.341	0.895	FAIL	1 colonne	CHK/CHKI
Rd_CG0_BL0_01	3.890	-2.000	0.896	PASS	-	ALL_00
Rd_CG0_BL0_1	0.780	- 1.018	0.895	FAIL	1 colonne/1 ligne	ALL_00
Rd_CG0_BL0_100	-0.412	0.125	0.895	FAIL	1 colonne/1 ligne	ALL_00
Capd_CG1_BL3_01fF	-2.340	- 2.340	0.895	FAIL	1ligne	ALL_11
…….	…….	…….	…….	…….	…….	…….

Tableau III. 13 Extrait de la base de données.

Toutes les informations relatives à chaque défaut simulé sont répertoriées dans le but de différencier les défauts les uns des autres.

E. Conclusion

Un mode de défaillance prédominant des circuits mémoires est dû à la présence des défauts aléatoires (particules isolées) lors de étapes de fabrication du circuit. Ces défauts aléatoires provoquent soit un court-circuit (« bridge ») soit un circuit ouvert (« break ») au niveau des lignes conductrices et peuvent être modélisés par l'ajout de résistances variables (0,1 kΩ à 1000 kΩ) au sein du circuit de simulation. De plus, l'augmentation de la densité d'intégration dans le plan mémoire provoque une augmentation du bruit de diaphonie entre interconnexions métalliques latérales. De ce fait, le couplage diaphonique devient un des phénomènes parasites majeur pour les technologies EEPROM avancées. Un dernier type de défaillance qui peut affecter le fonctionnement du plan mémoire est lié à l'activation de transistors parasites du fait de l'utilisation de signaux de programmation haute tension.
Une étude préalable de la technologie EEPROM (i.e. processus de fabrication) a été nécessaire de manière à mettre en évidence tous les défauts pouvant potentiellement affecter le fonctionnement du circuit.
A partir de cette étude, une procédure de détection systématique des causes de défaillances a été mise en place de manière à établir une corrélation entre les défauts simulés et les signatures électriques correspondantes.
Aux vues des résultats obtenus, il apparaît clairement que la mise en place d'une méthodologie de diagnostic passe par la connaissance de valeurs analogiques comme les tensions de seuil pour chaque cellule du plan mémoire.
Dans le chapitre IV, nous reviendrons plus en détail sur l'importance de l'extraction de données analogiques pertinentes représentatives des cellules mémoires EEPROM. Nous verrons que ces informations sont nécessaires à la mise en place d'une méthodologie de diagnostic de défauts efficace, en introduisant le concept de « bitmap » analogique.

CHAPITRE IV
Structures d'extraction embarquées des valeurs de seuil

I. Introduction

A. Etat de l'art

1. Signatures électriques dans les mémoires EEPROM

a. Problématique

Les mémoires EEPROM sont embarquées dans un grand nombre de circuits destinés à des applications spécifiques qui nécessitent des tailles mémoire différentes. Quelques bits de mémoires sont nécessaires à la personnalisation de cartes ou à la calibration de capteurs, alors que d'autres applications dédiées aux réseaux ou aux jeux vidéo nécessitent quelques mégabits de mémoire.
Le test des mémoires EEPROM est conditionné par la spécificité de ce type de mémoire dont les caractéristiques principales sont des temps de programmation longs et un caractère analogique lié au mécanisme de fonctionnement de la cellule mémoire (décalage des tensions de seuil représentatives de l'état électrique de la mémoire).
Les contraintes associées au nombre de broches externes sont imposées par l'application qui utilise la mémoire. Les mémoires à accès parallèle (stand alone memories) possèdent une observabilité et une contrôlabilité optimales, ce qui rend les tests et les étapes de caractérisation de la mémoire plus simples et plus rapides. Cependant, les contraintes économiques liées au surcoût engendré par un nombre de broches externes important tendent à généraliser l'utilisation des mémoires à accès série (cf. Chapitre III, §I.A.2a).
Dans le cas des systèmes sur puce et pour le cas particulier des cartes à puce, le test de la mémoire EEPROM est entièrement réalisé sur produit. La solution la plus utilisée consiste à générer les vecteurs de test à partir d'une mémoire ROM embarquée, préalablement codée. L'application des vecteurs de test est contrôlée par un microcontrôleur. Durant le test, les sorties de la mémoire EEPROM sont comparées aux valeurs attendues lors d'un mode de fonctionnement particulier de la carte, appelé mode de test. Si cette stratégie de test est optimale en production, le diagnostic de défauts au sein de la mémoire EEPROM est rendu très difficile de part la faible flexibilité de la méthode de test et de l'accès limité à la mémoire.
Les cartes à puce fabriquées sur le site de Rousset sont réalisées en technologie F8 0.18 µm et contiennent les éléments suivants (cf. Chapitre III, §I.A.2b) :

- un micro contrôleur,
- de la mémoire ROM associée au micro contrôleur,
- de la mémoire RAM associée au micro contrôleur,
- de la mémoire de données de type EEPROM.

Au niveau du processus de fabrication, ce type de produit nécessite plus de 300 opérations, 33 niveaux de masquage, l'utilisation de 4 oxydes différents (oxyde haute tension HV, basse tension LV, l'oxyde tunnel et l'Oxyde-Nitrure-Oxyde), 2 niveaux de polysilicium (grille de contrôle et grille flottante) et 4 niveaux de métaux. Le codage de la ROM étant réalisé au niveau du métal 1, le niveau de métal 4 étant un niveau de sécurité destiné à protéger l'accès à l'architecture du produit.
La complexité du processus de fabrication entraîne inévitablement l'apparition de défaillances que l'on peut associer aux différents éléments constituant la carte à puce.
Comme nous l'avons vu au chapitre précédent, les performances associées à la partie EEPROM sont très sensibles au processus de fabrication, de part la complexité du mécanisme

de fonctionnement de la cellule élémentaire et la surface occupée par le plan mémoire. De plus, une défaillance causée par un défaut dans une structure complexe comme un micro contrôleur est mise en évidence en un point du circuit généralement éloigné de la localisation physique du défaut. En contrepartie, un défaut dans la mémoire tend à causer une défaillance très proche de sa localisation physique, cela grâce aux caractéristiques de symétrie offertes par le plan mémoire [LEP94] [JEE93]. Le plan mémoire peut alors être utilisé comme un véhicule de test puissant de détection de défauts et de diagnostic des causes de ces défauts.
Les mémoires non volatiles de type EEPROM permettent l'extraction des distributions de tensions de seuil. Ces distributions ont la forme de gaussiennes et sont le résultat de la dispersion des paramètres technologiques de fabrication. Par conséquent, il est très important d'évaluer et de contrôler de manière précise les distributions des tensions de seuil des cellules pour chaque état électrique du plan mémoire. Et cela, de manière à mettre en évidence certaines populations de cellules marginales ou défaillantes. Plus encore, l'extraction de signatures électriques pertinentes au niveau d'une cellule mémoire EEPROM isolée sur produit, comme les tensions de seuil, pourrait apporter des informations utiles à une méthodologie de diagnostic de défauts, notamment durant les phases d'introduction de nouvelles technologies.

L'obtention de ces signatures électriques nécessite la mise en place de structures d'extraction embarquées sur le produit EEPROM. La conception et la mise en place de ces structures d'extraction sont au centre de ce chapitre. A partir de ces extractions, des signatures électriques pertinentes, obtenues sur produit, pourraient être utilisées par l'outil de diagnostic de défauts présenté au chapitre II, par exemple. Ce qui permettrait de remonter directement à la cause d'une défaillance (étape du processus de fabrication mise en cause) sans passer par de coûteuses études d'analyses de construction.

b. Répartition de l'effort de diagnostic dans l'EEPROM

Le diagnostic de défauts dans les mémoires de type EEPROM passe par une étude séparée de chaque élément constituant la mémoire comme le montre la figure IV.1. Ces éléments sont :

- la cellule mémoire élémentaire,
- la matrice de cellules,
- une partie analogique qu'on peut associer à la pompe de charge,
- une partie logique.

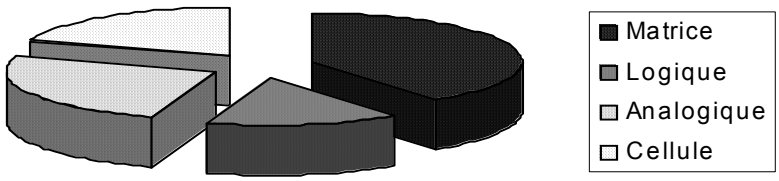


Figure IV. 1 Répartition de l'effort de diagnostic de défauts dans l'EEPROM.

Le chapitre II traite du diagnostic de défauts relatifs à la cellule mémoire EEPROM isolée, accessible sous forme de structure de test dans les lignes de découpe de la tranche de silicium.
Le chapitre III introduit une méthodologie de diagnostic de défauts dans le plan mémoire EEPROM.
Ces deux dernières méthodologies sont basées sur la connaissance de signatures électriques spécifiques aux mémoires EEPROM qui sont les tensions de seuil.

Le diagnostic de défauts relatifs aux parties logiques et analogiques ne sera pas traité.

2. Diagnostic de défauts et techniques de réparation

a. Rendement des mémoires EEPROM

La connaissance du rendement de produits complexes, comme ceux embarquant de la mémoire EEPROM, est de première importance puisqu'il conditionne la survie ou le maintien d'une technologie (conception et processus de fabrication). Il est donc important de disposer rapidement d'informations précises concernant les défaillances affectant, ou susceptibles d'affecter le produit EEPROM. Dans le cas des mémoires non volatiles, des études statistiques ont montré que la baisse du rendement était principalement due à des défauts affectant le plan mémoire (ce qui semble évident au vu de la surface importante occupée par la partie mémoire au sein d'une puce).
Les défaillances les plus fréquentes qui affectent la matrice mémoire sont de deux types :

- courts-circuits au niveau des pistes métalliques affectant deux lignes (parfois deux colonnes) adjacentes,
- cellules mémoires dont le comportement électrique est incorrect.

Le modèle statistique le plus répandu pour décrire le rendement utilise l'approximation de « Poisson » dont l'expression prend la forme suivante :

$$Y_{tot} = e^{(-AD)} \quad \text{(IV. 1)}$$

Y_{tot} représente le rendement total, A la surface de la puce considérée, et D le nombre de défauts par unité de surface. Ce rendement peut être décomposé en deux rendements élémentaires Y_m et Y_p, qui se réfèrent respectivement à la partie mémoire et à la partie regroupant les circuits périphériques. Cette étape est nécessaire au calcul du rendement puisqu'il existe des techniques de réparation de défauts (comme l'analyse de redondance) qui touchent uniquement le plan mémoire. En nommant A_m et A_p l'aire des deux parties de la puce, et D_m et D_p la densité de défauts correspondante, nous pouvons écrire :

$$A_m.D_m + A_p.D_p = AD \quad \text{(IV. 2)}$$

En substituant cette relation dans l'équation IV.1, on a :

$$Y_{tot} = Y_m.Y_p = e^{(-ApDp)}e^{(-AmDm)} \quad \text{(IV. 3)}$$

A ce stade, il est opportun d'introduire les deux paramètres suivants :

$$k = \frac{A_p}{A_m}, \quad R = \frac{D_p}{D_m} \quad \text{(IV. 4)}$$

Ce qui permet d'exprimer Y_m et Y_p de la manière suivante :

$$Y_p(Y_{tot}, k, R) = Y_{tot}^{1-\frac{1}{1+kR}} \quad \text{(IV. 5)}$$

$$Y_m(Y_{tot}, k, R) = Y_{tot}^{\frac{1}{1+kR}} \quad \text{(IV. 6)}$$

Le rendement Y_m nous intéresse particulièrement puisqu'il peut être corrélé à la probabilité p qu'a une cellule mémoire d'être défaillante. En assumant que les défaillances de cellules sont indépendantes dans le plan mémoire et que toutes les défaillances sont associées à des cellules mémoires isolées, le rendement relatif à la partie matrice peut alors s'écrire :

$$Y_m = (1-p)^{N_{cellules}} \quad \text{(IV. 7)}$$

$$p(Y_m, N_{cellules}) = 1 - Y_m^{\frac{1}{N_{cellules}}} \quad \text{(IV. 8)}$$

Pour être complet, il est nécessaire de rappeler la spécificité de la partie matrice mémoire EEPROM par rapport aux circuits périphériques. D'une manière générale, les oxydes minces sont les éléments les plus critiques pour le fonctionnement des circuits VLSI. Dans le cas des mémoires non volatiles de type EEPROM, l'oxyde tunnel est de loin l'élément le plus sensible du processus de fabrication. De plus, l'augmentation des densités d'intégration a considérablement augmenté la surface totale occupée par l'oxyde tunnel.
A cela s'ajoute le « stress » important auquel est soumis le plan mémoire (10 MV/cm aux bornes de l'oxyde tunnel, durant les étapes de programmation) comparativement à la partie logique du produit EEPROM (5 MV/cm au maximum pour la partie haute tension). De plus, contrairement aux oxydes de la partie logique dont le rôle est de garantir une bonne isolation, les oxydes des cellules du plan mémoire doivent offrir de bonnes performances en rétention, c'est-à-dire conserver l'information stockée pendant plusieurs années.
A partir de ces constatations, des techniques dites de réparation ont été mises en œuvre de manière à assurer une meilleure fiabilité du plan mémoire EEPROM et maintenir un rendement de fabrication élevé.

b. Analyse de redondance

Le rendement des mémoires non volatiles est généralement lié à deux types de défauts : les défauts paramétriques et les défauts aléatoires. Les défauts paramétriques sont associés à des variations du processus de fabrication, mais aussi aux méthodes de conception utilisées. Ces types de défauts affectent de nombreuses puces sur une tranche de silicium et entraînent par conséquent un effet catastrophique sur le rendement, ce qui permet de les identifier et de les corriger rapidement.
Les défauts aléatoires sont le résultat d'une variation localisée (lacune, surplus ou déplacement) d'un matériau constituant le circuit. Ce qui peut entraîner une défaillance si ces défauts sont localisés en un point particulier du circuit. Ce type de défauts peut affecter n'importe quel niveau de masque ou survenir à n'importe quelle étape du processus de fabrication. Il entraîne, dans la plupart des cas, l'apparition de colonnes défaillantes dans le plan mémoire. Cette défaillance inclut des contacts de drain ou de source manquants ou encore des niveaux de métaux de « bit line » ouverts ou court-circuités. Pour remédier à ce dernier type de défauts, des techniques de redondance sont utilisées de manière à améliorer le rendement. La circuiterie de redondance permet de substituer les colonnes ou lignes défaillantes d'un plan mémoire par des colonnes ou lignes de « rechange », généralement

situées sur les côtés de la mémoire. Le nombre d'éléments redondants nécessaires à un réajustement efficace du rendement prend en compte plusieurs paramètres dont :

- la surface utilisée par les éléments de redondance,
- la taille du plan mémoire,
- la taille des défauts susceptibles d'altérer le plan mémoire [SEG99],
- les géométries minimales de la technologie considérée [SEG99],
- la surface totale de la puce qui embarque la mémoire,
- la qualité du processus de fabrication.

La détection et l'implémentation des techniques de redondance sont réalisées durant le premier flot de test. Le processus de réparation comprend trois étapes :

- la détection des fautes durant le flot de test et la mémorisation des adresses physiques des cellules mémoires défaillantes,
- une étape de diagnostic où l'analyse des erreurs sauvegardées dans la mémoire du testeur va permettre de déterminer si la puce est réparable, en fonction des éléments de redondance disponibles,
- l'activation des ressources de réparation.

Les testeurs utilisés doivent disposer de ressources spécifiques de manière à pouvoir fournir une représentation topologique de la mémoire, différenciant toutes les cellules mémoires testées (ligne, colonne et résultat de test de la cellule). La stratégie de réparation doit être mise en place après que tous les tests fonctionnels aient été effectués, c'est-à-dire après l'obtention qu'une cartographie précise et définitive de toutes les cellules défaillantes du plan mémoire.
L'activation et l'accès à chaque ressource de réparation sont contrôlés par une circuiterie spécifique appelée CAM (Content Adressable Memory), parfaitement intégrée dans l'architecture de la mémoire [JEX91].
Un modèle simplifié d'implémentation de la redondance est donné figure IV.2. L'adresse d'une ligne défaillante par exemple, est dans un premier temps mémorisée dans la CAM après le premier flot de test. Ensuite, à chaque fois que la ligne défaillante sera sélectionnée, la CAM va désélectionner le décodeur de lignes de la mémoire principale et se substituer à ce dernier de manière à adresser directement une ligne de rechange. La sauvegarde définitive des adresses des colonnes ou des lignes défaillantes utilise de la mémoire non volatile, qui a l'avantage de convenir au processus de fabrication et de se tester facilement. Cette mémoire se situe à l'intérieur de la CAM.
Le gain obtenu en terme de rendement peut être considérable. Cependant, cette technique reste difficile à mettre en œuvre et a un impact direct sur la surface totale occupée par la puce ainsi que sur le temps d'accès mémoire. De plus, le temps de test ainsi que la complexité du circuit augmente. La mise en place de techniques de redondance doit donc être un compromis entre la taille de la puce, les quantités produites et le coût du test [DAG02].
Il est à noter que cette technique de réparation ne s'applique qu'à la partie mémoire, et non aux circuits périphériques. En effet, la partie mémoire possède une structure répétitive et utilise les règles de conception minimales pour la technologie considérée. Alors que les circuits périphériques sont des éléments hétérogènes et n'utilisent pas systématiquement les géométries minimales. Il en résulte que l'utilisation de la redondance pour les circuits périphériques ne serait pas économiquement viable du fait du nombre important d'éléments redondants à mettre en place sur la puce. De plus, la probabilité de défaillance associée aux circuits périphériques est faible en comparaison de la partie mémoire.

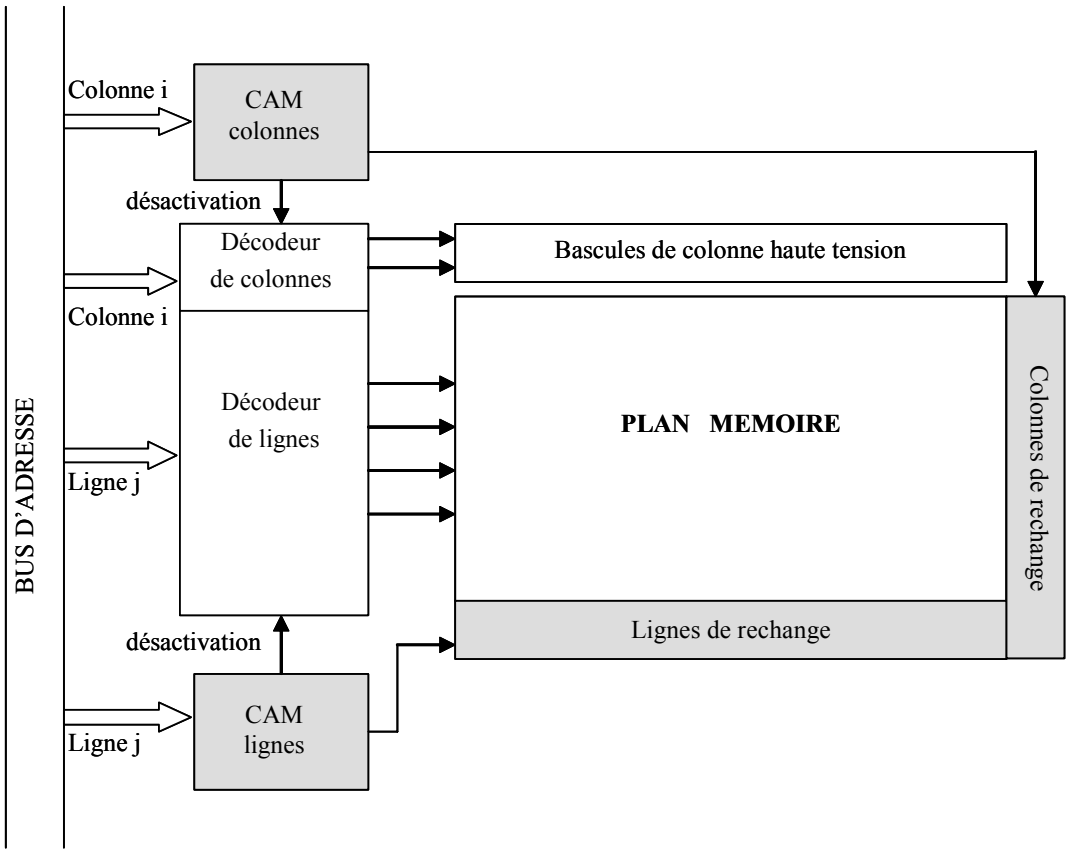


Figure IV. 2 Modèle fonctionnel simplifié de redondance.

c. Codes de correction d'erreurs

D'autres techniques connues sous le nom de codes de correction d'erreurs (Error Correction Code ou plus simplement ECC) [WAL94] sont utilisées de manière à améliorer la fiabilité de la mémoire. Bien qu'encore peu utilisés, les codes de correction d'erreurs sont une alternative à la redondance. Cependant, bien que l'on puisse considérer qu'il s'agit d'un moyen d'augmenter le rendement au sein d'une puce mémoire, l'utilisation de cette technique est le plus souvent dictée par des impératifs de fiabilité plutôt que par une augmentation des rendements.

Les codes de correction d'erreurs sont une extension des codes de détection d'erreurs (Error Detection Code ou plus simplement EDC). L'exemple d'implémentation de codes de détection d'erreurs le plus simple et le plus utilisé est l'utilisation du bit de parité. Lorsqu'un mot est écrit en mémoire, un bit de parité est généré à partir des bits composants le mot mémoire. Lors d'une opération de lecture, un calcul du bit de parité est effectué sur le mot lu, et le résultat de calcul est comparé au bit de parité du mot écrit en mémoire. Une différence entre les bits de parité signifie qu'il existe au moins une erreur pour le mot considéré (ou une erreur au niveau du bit de parité lui-même). Cependant, l'utilisation d'un seul bit de parité ne permet de détecter qu'un nombre impair d'erreurs et, de ce fait, un nombre pair d'erreurs restera indétectable. Malgré cette limitation, la majorité des erreurs qui affectent un mot binaire sont dues à la défaillance d'un seul bit et sont toujours détectées.

Les codes de correction d'erreurs permettent, quand à eux, de détecter puis de corriger les erreurs affectant un mot binaire. Les systèmes de correction d'erreurs ajoutent des bits supplémentaires ou redondants à un mot binaire. Ces bits redondants procurent une certaine structure au mot binaire. Si cette structure venait à être altérée ou modifiée suite à des défauts

physiques ou au vieillissement de la puce, ces changements seraient détectés et corrigés. Dans le cas des mémoires non volatiles de type EEPROM, les erreurs doivent être détectées et corrigées instantanément, ce qui implique une implémentation physique sur puce de la logique de détection et de correction d'erreurs. Le premier ouvrage consacré aux codes de correction d'erreur a été écrit en 1961 par W. Wesley Peterson [PET61].
Les algorithmes de détection et de correction d'erreurs s'appliquent à des bits ou des mots binaires. D'un point de vue mathématique chaque bit ou mot binaire représente un élément dans un espace fini, ce qui correspond à un système algébrique avec un nombre fini d'éléments, deux opérations et leurs inverses. Ces opérations sont la multiplication, l'addition et leurs inverses sont la division et la soustraction. De ce fait, il est simple d'appliquer la théorie des éléments finis à des circuits mémoires ou digitaux [BER68].

d. Spécificité des mémoires non volatiles

L'utilisation de la redondance ou des codes de correction d'erreurs est souvent conditionnée par des contraintes liées à l'application qui utilise la mémoire. Cette utilisation induit systématiquement un surcoût en terme de surface silicium et de temps de test.
D'autres méthodes visant à une amélioration du rendement sont basées sur une cartographie bit ou « bitmap » de la mémoire, telle que celle proposée dans l'article de J. Segal [SEG01]. Ces méthodes ont été utilisées avec succès dans le cas des mémoires non volatiles. Toutefois, elles ne prennent pas en compte le mécanisme analogique de mémorisation, à savoir le basculement des tensions de seuil de la cellule EEPROM.
En effet, je rappelle que dans le cas d'une mémoire EEPROM de technologie dite « Flotox », maintenant devenue la norme dans l'industrie, le transfert de charge électrique depuis le nœud drain vers le nœud de grille flottante est dû à l'effet tunnel Fowler-Nordheim. Ce courant d'injection tunnel (décrit en détail dans chapitre I, §I.B.1a) entraîne un décalage des tensions de seuil et dépend entièrement des facteurs suivants :

- la géométrie de la cellule,
- la haute tension appliquée pour programmer la cellule,
- la qualité du processus de fabrication.

L'ensemble de ces éléments mène à des problèmes critiques de conception et de contrôle de processus industriels dans le domaine des mémoires non volatiles. Une variation, même faible, de l'un de ces facteurs parmi les cellules d'un plan mémoire, a pour effet direct un étalement de la distribution des tensions de seuil, ce qui se traduit par une diminution à la fois des performances globales de la mémoire et du rendement de fabrication. Il apparaît donc important de pouvoir disposer de techniques de diagnostic spécifiques au produit EEPROM, basées sur l'obtention des tensions de seuil de chaque cellule du plan mémoire.

3. Le bitmap « logique »

a. Flot de test fonctionnel standard

Le test fonctionnel des mémoires EEPROM est soumis à plus de contraintes que celui des mémoires vives (RAM) classiques. En effet, l'opération de programmation est bien trop coûteuse en temps de test pour permettre l'utilisation d'algorithmes tels que ceux de type March (cf. Chapitre I, §III.A.6). Ainsi, on préfère adopter pour les mémoires EEPROM des procédures de test de type n tels que :

- programmation de toutes les cellules à l'état logique 1,
- programmation de toutes les cellules à l'état logique 0,
- programmation d'un « damier »,
- programmation d'une ligne ou d'une colonne,
- programmation d'une « diagonale ».

Plus d'informations sur ces techniques sont données dans le livre de M. Cappelletti [CAP99]. Tous ces algorithmes de test simples ne requièrent qu'un cycle de programmation (effacement puis écriture) suivie d'une opération de lecture.
Par ailleurs, il est important de retenir que le temps de lecture représente moins de 0,1% du temps total de programmation.

b. Limitations du bitmap logique

En terme de diagnostic, une cartographie de bit logique (« bitmap ») est réalisée pour chacun des algorithmes de test appliqués aux mémoires non volatiles de type EEPROM. La disposition particulière des cellules défaillantes du plan mémoire fournit des signatures électriques à partir desquelles il est possible dans certains cas, de remonter à l'origine physique de la défaillance. Ces signatures électriques peuvent représenter des lignes, des colonnes, des groupes de bits ou des bits isolés comme le montre la figure IV.3.

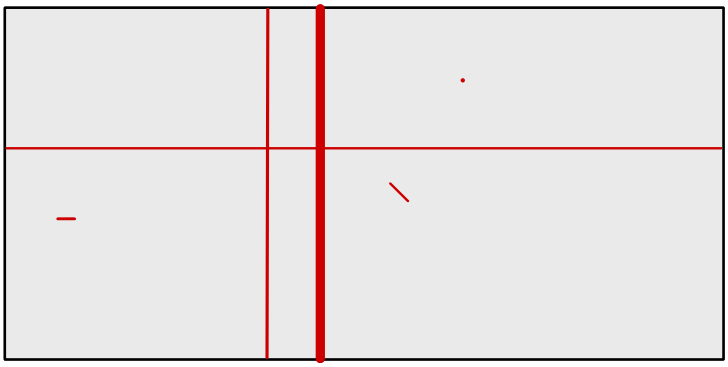


Figure IV. 3 « bitmap ».

Les limitations de ce type de représentation sont évidentes :

- la bibliothèque de signatures obtenue est limitée,
- des défaillances différentes peuvent conduire à des signatures électriques identiques,
- les tensions de seuils hors spécifications ne produisent aucune signature.

Cette représentation « bitmap » est obtenue à la suite de l'exécution du flot de tests fonctionnels standard des mémoires EEPROM. Ces tests ne fournissent aucune information de nature analogique, alors que les mémoires EEPROM présentent un mécanisme de mémorisation de type purement analogique.
En effet, comme le montre la figure IV.3, la seule information disponible pour une cellule défaillante est sa localisation dans le plan mémoire.

B. Conception en vue de l'analyse de défaillance

1. Extraction embarquée de signatures électriques

Une méthodologie dédiée au diagnostic de défauts dans les mémoires EEPROM passe par la mise en place de moyens intégrés d'extraction de signatures électriques comme les tensions de seuil, mais aussi par une modification de la séquence de test classique. Grâce à la mise en place de ces moyens supplémentaires, on va pouvoir extraire rapidement et avec précision les tensions de seuil de toutes les cellules d'une mémoire EEPROM. La méthodologie de diagnostic classique est alors modifiée pour prendre en compte les valeurs de seuil de chaque cellule mémoire.
L'importance de l'évaluation et du contrôle précis des tensions de seuil est rendue nécessaire face à l'augmentation des densités d'intégration amenée par une diminution drastique des règles de dessin. Ceci a conduit à un élargissement des distributions des tensions de seuil du plan mémoire EEPROM, particulièrement pour les nouvelles technologies [SAK94]. De plus, de manière à réaliser des mémoires embarquées de faible consommation destinées à servir le marché des dispositifs portatifs telles que les cartes à puce, il est nécessaire de connaître les tensions de seuil de chaque cellule du plan mémoire et de contenir l'élargissement de leurs distributions.
Des techniques d'évaluation des tensions de seuil, basées sur des structures de test ont été proposées ces dernières années par T.Himeno [HIM95] et K.Hakozaki [HAK97] pour les mémoires Flash EEPROM. L'une d'entre elles utilise des circuits périphériques additionnels spécifiques pour évaluer les distributions des tensions de seuil. Une autre technique permet d'évaluer les populations de cellules mémoires marginales en mesurant le courant sous le seuil de la totalité de la matrice de cellules constituant le plan mémoire. Une innovation à ces techniques classiques a été présentée par R. Khubchandani [KHU97] qui introduit une représentation topologique des distributions des tensions de seuil sur tout un plan mémoire, sous forme de « bitmap » coloré. Pour cela, une structure de test spécifique qui permet un accès direct aux cellules mémoire (mode de test DMA) a été utilisée. Le point commun à toutes ces techniques est bien entendu l'utilisation de structures de test spécifiques qui accompagnent le produit embarquant la mémoire EEPROM.
L'approche utilisée va consister à mettre en place des structures d'extraction des valeurs de seuil directement sur le produit de manière à disposer de données analogiques fiables relatives aux cellules mémoires du produit EEPROM, sans passer par des véhicules de test. Cette approche va nous permettre de nous orienter vers une méthodologie de diagnostic basée sur des moyens intégrés d'extraction.

2. Limitations dans le cas du produit EEPROM

a. Extraction du courant de seuil

Dans les produits EEPROM standard, la tension de seuil de la cellule dans l'état électrique écrit ne peut pas être directement mesurée. Cela est dû à la circuiterie de décodage des lignes de la matrice mémoire, qui n'est pas conçue pour transférer des tensions négatives.

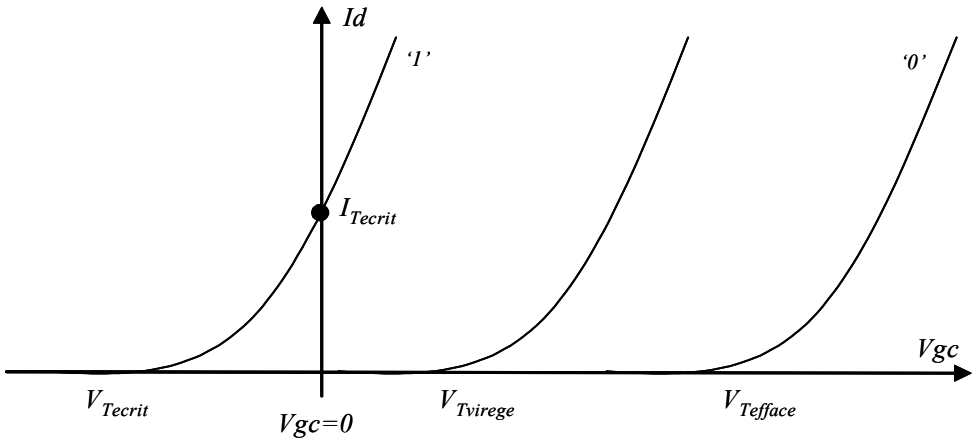


Figure IV. 4 Tensions de seuil représentatives de l'état de la cellule EEPROM.

Cependant, les valeurs des tensions de seuil dans l'état vierge et effacé peuvent être extraites de manière classique. Concernant l'état écrit de la cellule mémoire, une approche alternative va consister à extraire le courant de seuil I_{Tecrit} obtenu pour une tension de grille V_{gc} nulle comme le montre la figure IV.4. Ce courant de seuil est l'image de la tension de seuil V_{Tecrit} sur la caractéristique I_d (V_{gc}) du transistor mémoire. Comme nous le verrons par la suite, cette extraction du courant de seuil se fera en augmentant par palier le courant de lecture I_d jusqu'à ce que le transistor mémoire devienne passant.

b. Contraintes associées aux structures d'extraction embarquées

La mise en œuvre pratique de l'extraction des tensions de seuil, soit directement, soit par le biais d'une mesure de courant de seuil, passe par la réalisation de structures d'extraction intégrées (« Built In Self Diagnosis ») sur produit. Les contraintes liées à ces structures sont de trois ordres :

- la surface occupée sur la puce par les structures d'extraction doit être la plus faible possible,
- le surcoût en terme de temps de test doit être limité autant que possible,
- le mode de fonctionnement normal de la mémoire ne doit pas être affecté.

La valeur de la tension de seuil ou du courant de seuil doit être fournie après la phase de test de la mémoire au format numérique (digital). Cela, de manière à ce que ces valeurs de seuil soient compatibles ou facilement observables à l'aide d'une logique aléatoire dans le cas des mémoires EEPROM embarquées.
Bien entendu, la mise en place sur produit des structures d'extraction sera accompagnée d'une modification du flot de test standard de manière à exploiter ces structures d'extraction embarquées et obtenir de nouvelles informations de type analogique.
Plus précisément, la modification du flot de test standard portera uniquement sur la phase de lecture de la cellule mémoire qui, je le rappelle, est très peu consommatrice en terme de temps de test.

II. Extraction des signatures électriques

A. Extraction embarquée des tensions de seuil

1. La tension de seuil du transistor mémoire

a. Définition de la tension de seuil

La tension de seuil peut être définie comme la tension minimale à appliquer sur le transistor mémoire de manière à créer un courant de conduction entre la source et le drain du transistor. Plus exactement, cette définition est utilisée en spécifiant un courant de conduction donné nommé I_{lec}. Par exemple, S.Kobayashi [KOB95] définit la tension de seuil V_{th} comme la tension de grille qui permet au courant de drain d'atteindre une valeur de 10 µA. Une définition similaire est donnée par D.Wellekens [WEL95]. Dans ce cas, la tension de seuil est donnée par une tension de grille qui correspond à un courant de drain de 1 µA pour une tension de drain égale à 2 V et une tension de grille du transistor de sélection égale à 3 V.
Dans tous les cas et pour une définition donnée, la valeur de la tension de seuil reste proportionnelle à la charge de la grille flottante.

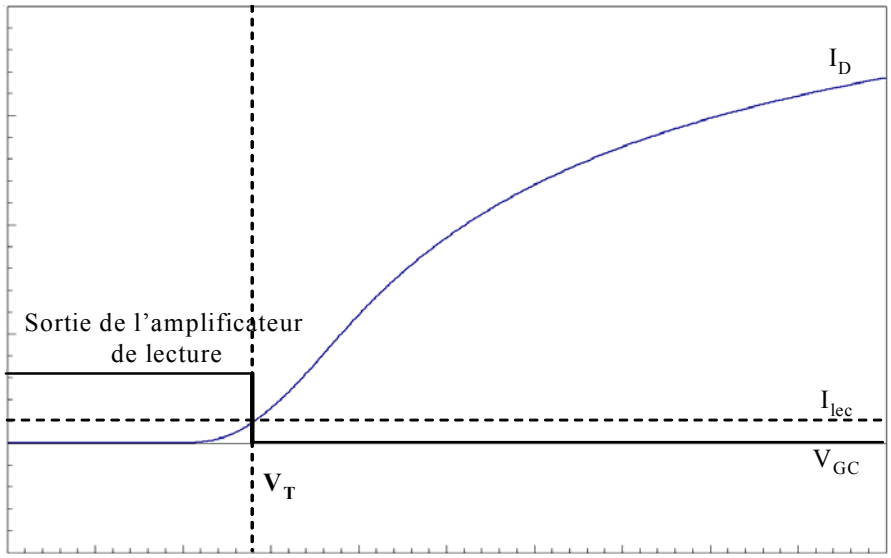


Figure IV. 5 Définition de la tension de seuil.

La valeur de la tension de seuil d'une cellule mémoire doit être définie dans les conditions du composant (c'est-à-dire au sein de la matrice mémoire embarquée et non pas en utilisant des structures de test élémentaires). La figure IV.5 représente l'évolution du courant I_D traversant la cellule mémoire en fonction de la tension de grille V_{GC} pour une tension de drain V_{DS} fixée. Du point de vue produit, la valeur de la tension de seuil est donnée par la valeur de la tension de la grille de commande V_{GC} lorsque la sortie de l'amplificateur de lecture commute. Au niveau électrique, on peut noter que la valeur de la tension de seuil V_T est donnée par l'intersection du courant de lecture I_{lec} et de la caractéristique du courant de drain I_D.
Dans le cadre de notre étude qui concerne des mémoires EEPROM fabriquées par la société STMicroelectronics (Rousset) à partir de la technologie F8 0,18 µm, le courant de lecture I_{lec} est fixé à 10 µA pour un tension de drain égale à 0,8 V.

b. Distribution des tensions de seuil

D'une manière générale l'extraction des distributions des tensions de seuil permet d'évaluer le comportement d'un grand nombre de transistors du plan mémoire durant son fonctionnement. Pour chaque état électrique du plan mémoire, les valeurs des tensions de seuil de toutes les cellules mémoire sont extraites de manière à construire les distributions des tensions de seuil, comme le montre la figure IV.6.
Ces distributions sont obtenues en faisant varier la tension de grille de contrôle V_{GC}. Pour chaque pas de variation de la tension de grille, et pour chaque cellule (ou mot) adressée, les commutations de l'amplificateur de lecture permettent de comptabiliser le nombre de cellules dont la tension de seuil a été atteinte (i.e. tension V_{GC} qui provoque la commutation de l'amplificateur de lecture). Cette méthode permet de ramener le calcul des distributions des tensions de seuil à une succession d'opérations de lecture [CAP99].

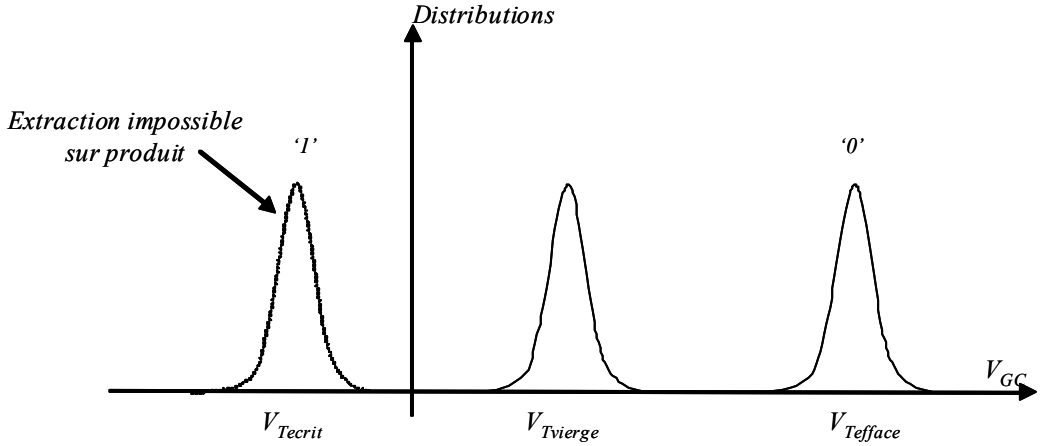


Figure IV. 6 Distribution des tensions de seuil d'un plan mémoire.

De même que pour l'extraction de la tension de seuil de la cellule dans l'état électrique écrit, la distribution des cellules écrites n'est pas mesurable sur produit.

2. Principe d'extraction de la tension de seuil

a. Lecture classique d'une cellule mémoire

Le résultat de lecture de la cellule mémoire EEPROM dépend de la valeur de la tension de grille de contrôle V_{GC}. A partir de la courbe de la figure IV.5, les observations suivantes définissent l'état de sortie de l'amplificateur de lecture en fonction de la tension V_{GC} :

- lorsque la tension V_{GC} de la grille de commande est en dessous de la tension de seuil V_T pour un courant de lecture I_{lec} donné, la cellule n'est pas capable de conduire le courant de lecture, et la sortie de l'amplificateur présente un état haut (logique 1),
- lorsque la tension V_{GC} est au-dessus de la tension de seuil V_T pour un courant de lecture I_{lec} donné, la cellule laisse passer tout le courant de lecture, et la sortie de l'amplificateur de lecture passe alors à un état bas (logique 0).

Lors d'une opération de lecture classique, un potentiel prédéterminé V_{CGref} (dont la valeur se situe au niveau de la distribution des cellules dans l'état vierge) est appliqué sur la grille de commande de la cellule adressée. L'amplificateur de lecture applique ensuite le courant de

lecture I_{lec} sur la ligne ou « bitline » sélectionnée. Si la cellule lue est à l'état écrit, le courant de lecture I_{lec} parcourt la cellule sélectionnée, ce qui entraîne une commutation de l'amplificateur de lecture. Si la cellule est dans l'état effacé, le transistor à grille flottante est bloqué et aucun courant ne traverse la cellule mémoire. Dans ce cas, il ne se produit aucune commutation au niveau de la sortie de l'amplificateur de lecture. Le fonctionnement de l'amplificateur de lecture est rappelé à la section I.B du chapitre I.
Cette lecture, dite classique, est utilisée dans le flot de test standard des mémoires EEPROM. Elle permet de détecter si une cellule est écrite ou effacée. Mais ce type de lecture ne fournit en aucun cas des informations de type analogique relatives à la cellule adressée.

b. Lecture modifiée de la cellule EEPROM

Une procédure de lecture modifiée du point mémoire EEPROM devra permettre l'extraction des tensions de seuil sur le produit EEPROM. Cette extraction devra se faire en intégrant une structure spécifique dans le produit EEPROM.
Contrairement à une lecture classique du point mémoire où le potentiel de grille reste fixe, la tension de grille sera ici variable et la précision de la valeur de la tension de seuil extraite dépendra de la finesse du pas de variation. La tension de seuil étant obtenue pour une tension de grille qui rendra le transistor mémoire conducteur.

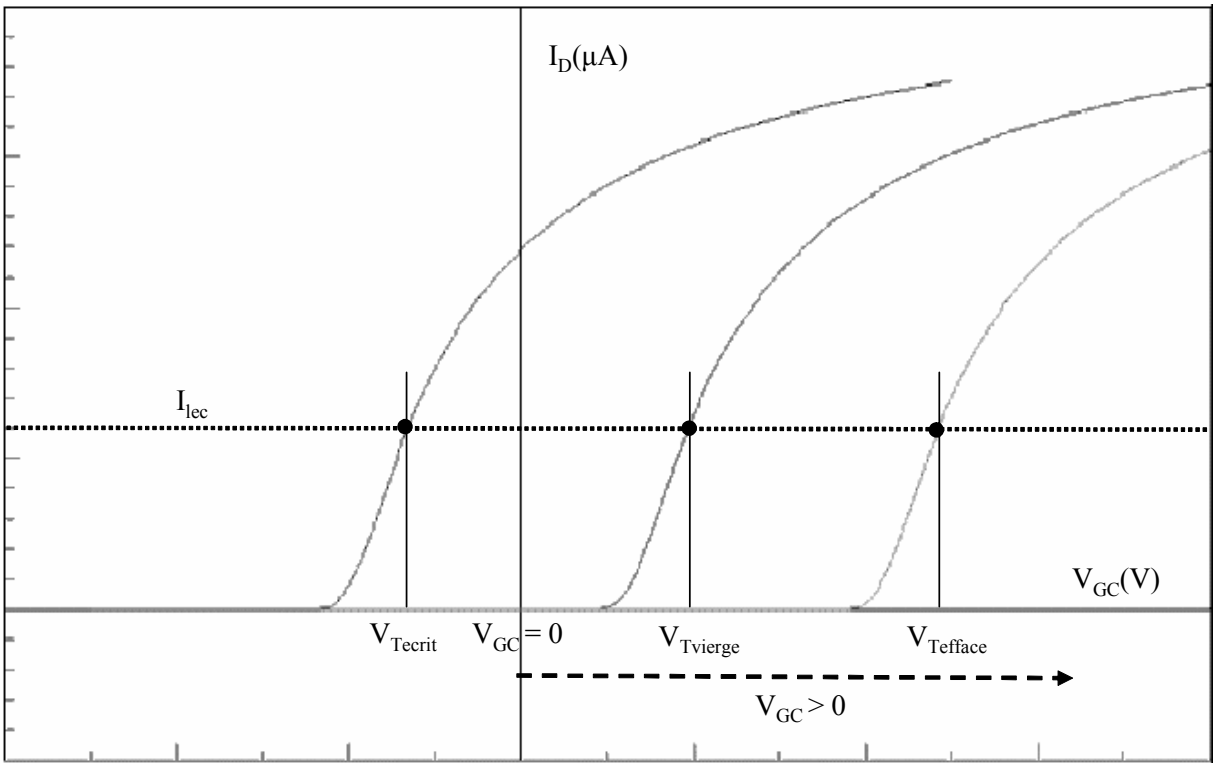


Figure IV. 7 Principe d'extraction des tensions de seuil vierges et écrites.

La seule limitation de cette technique concerne l'extraction de la tension de seuil V_{Tecrit} (obtenue pour $V_{GC} < 0$) de la cellule mémoire. En effet, comme nous l'avons vu, pour extraire la tension de seuil écrite il faudrait être capable d'appliquer une tension négative variable sur la grille du transistor mémoire, ce qui n'est pas réalisable sur produit. Cependant, l'extraction des tensions de seuil vierges et effacées (obtenues pour $V_{GC} > 0$) peut être envisagée sur produit comme l'indique la figure IV.7.

3. Structures d'extraction embarquées de la tension de seuil

a. Structure d'extraction : génération d'une tension de grille variable

Les conditions d'extraction de tensions de seuil $V_{Tvierge}$ et $V_{Tefface}$ sont données par les conditions de lecture suivantes :

- le courant injecté à partir de l'amplificateur de lecture est fixé à sa valeur nominale I_{lec},
- la tension de grille V_{GC} augmente par pas de valeur V_{REF}.

La génération d'une tension de grille variable est réalisée en utilisant des références de tension élémentaires [UYE97]. Une représentation schématique d'une partie de ce circuit est donnée figure IV.8. Le module d'extraction de la tension de seuil se compose de sources de tensions élémentaires montées en parallèle. Ces sources de tension sont contrôlées par des portes de transmission. Une sélection successive de chaque source de tension élémentaire permet d'obtenir un signal évoluant par palier de valeur V_{REF} au niveau du nœud de sortie de la structure V_{GC}.

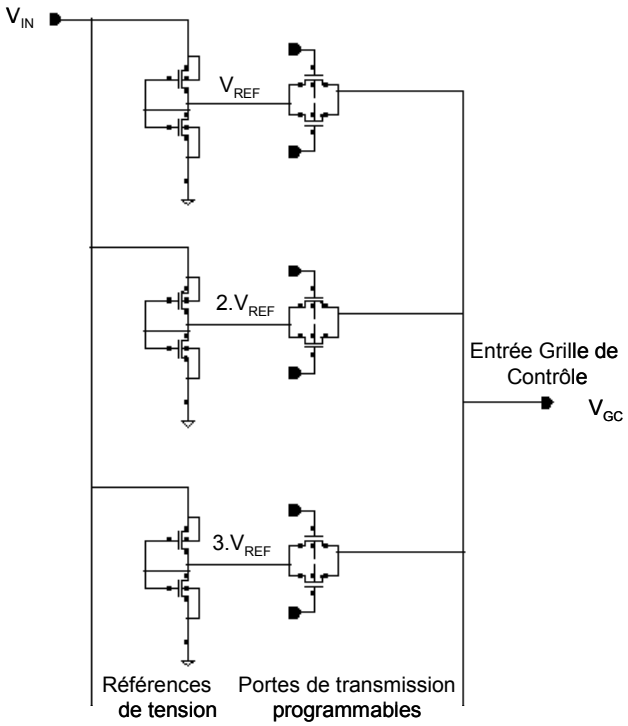


Figure IV. 8 Source de tension programmable par palier.

L'intégration de cette structure au sein du produit EEPROM permettra de fournir une tension de grille de contrôle V_{GC} variable pour chaque cellule sélectionnée du plan mémoire. La structure d'extraction de la figure IV.8 est en réalité composée de 20 étages, ce qui permettra de fournir une tension de sortie maximale égale à $20V_{REF}$.

b. Procédure d'extraction

Les portes de transmission sont contrôlées par un registre à décalage de manière à fournir sur la sortie V_{GC} un signal « en escalier » de pas V_{REF}. Le registre à décalage utilisé comporte 20 bits, il est initialisé avec tous ses bits à l'état logique '0', sauf le premier qui se trouve à l'état

logique '1'. La figure IV.9 présente un exemple d'extraction de la tension de seuil. L'extraction de la tension de seuil commence par l'application d'une tension de valeur V_{REF} sur la grille du transistor mémoire et le processus de décalage du bit '1' dans le registre prend fin au moment où l'amplificateur de lecture commute (la détection de la tension de seuil provoque une commutation de l'amplificateur de lecture et ce signal met fin au processus de décalage). A ce moment précis, la valeur de la tension de seuil V_T est mémorisée au format numérique et peut être obtenue à partir de la valeur binaire contenue dans le registre à décalage.

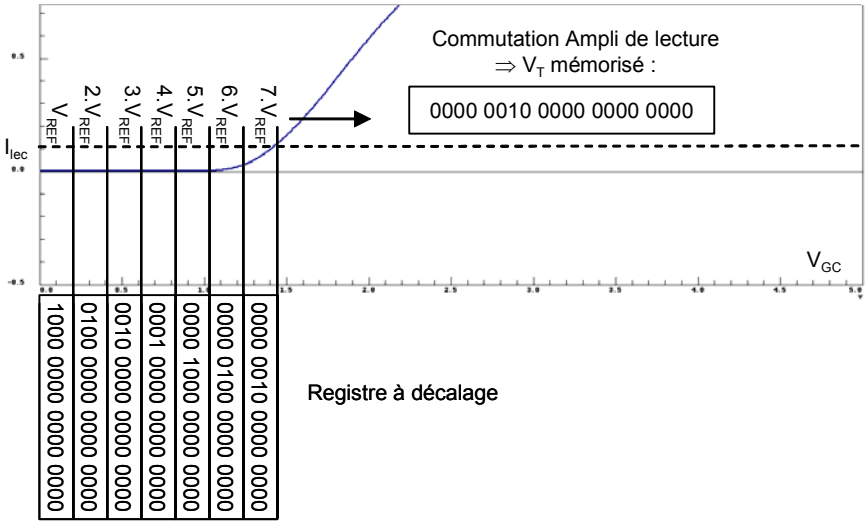


Figure IV. 9 Procédure d'extraction de la tension de seuil V_T.

Un dimensionnement précis des références de tension élémentaires sera nécessaire de manière à obtenir un pas de variation constant de la tension de sortie V_{GC} de la structure. De plus, il est à noter que la même structure est utilisée pour l'extraction des valeurs de seuil $V_{Tvierge}$ et $V_{Tefface}$. Le seul changement étant lié à la tension d'entrée V_{in} de la structure qui est portée à deux potentiels différents :

- V_{boost}, dans le cas de l'extraction de la tension de seuil $V_{Tefface}$, V_{boost} est fournie par un circuit multiplicateur de tension ou pompe de charge [TAN99];
- V_{dd}, pour l'extraction de la tension de seuil $V_{Tvierge}$.

B. Extraction embarquée des courants de seuil

1. Le courant de seuil du transistor mémoire

a. Définition du courant de seuil

A partir de la courbe de la figure IV.10, les observations suivantes permettent de définir le courant de seuil :

- lorsque le courant de lecture est en dessous du courant de seuil I_T, pour une tension de lecture V_{GClec} donnée, la cellule mémoire laisse passer le courant de lecture appliqué, et la sortie de l'amplificateur présente un état haut.
- lorsque le courant de lecture est au-dessus du courant de seuil I_T pour une tension de lecture V_{GClec} donnée, la cellule mémoire n'est pas capable de

conduire le courant de lecture appliqué, et la sortie de l'amplificateur de lecture passe alors à l'état bas.

L'extraction du courant de seuil est réalisée pour une tension grille V_{GClec} fixe. Ensuite, le courant de lecture augmente progressivement jusqu'à ce qu'il atteigne le courant de seuil I_T. A ce moment, le transistor mémoire devient passant, ce qui provoque une commutation de l'amplificateur de lecture.
De la même manière que les distributions des tensions de seuil, les distributions des courants de seuil peuvent être obtenues en adressant les cellules d'un plan mémoire une par une et en mémorisant leurs courants de seuil respectifs après extraction.

b. Spécifications en courant

Il est important de garder en tête que le courant de seuil n'est qu'une image de la tension de seuil sur la caractéristique $I_D(V_{GC})$. Il fournit donc, par définition, les mêmes informations et peut être utilisé de la même manière pour diagnostiquer les défauts dans l'EEPROM.
La technique d'extraction du courant de seuil passe par la définition d'une fenêtre de mesure dans laquelle variera le courant de lecture. La figure IV.10 montre que la fenêtre de spécification de la tension de seuil devient une fenêtre de spécification en courant de seuil. Cette correspondance est obtenue en utilisant la caractéristique de transfert $I_D(V_{GC})$ du transistor mémoire.

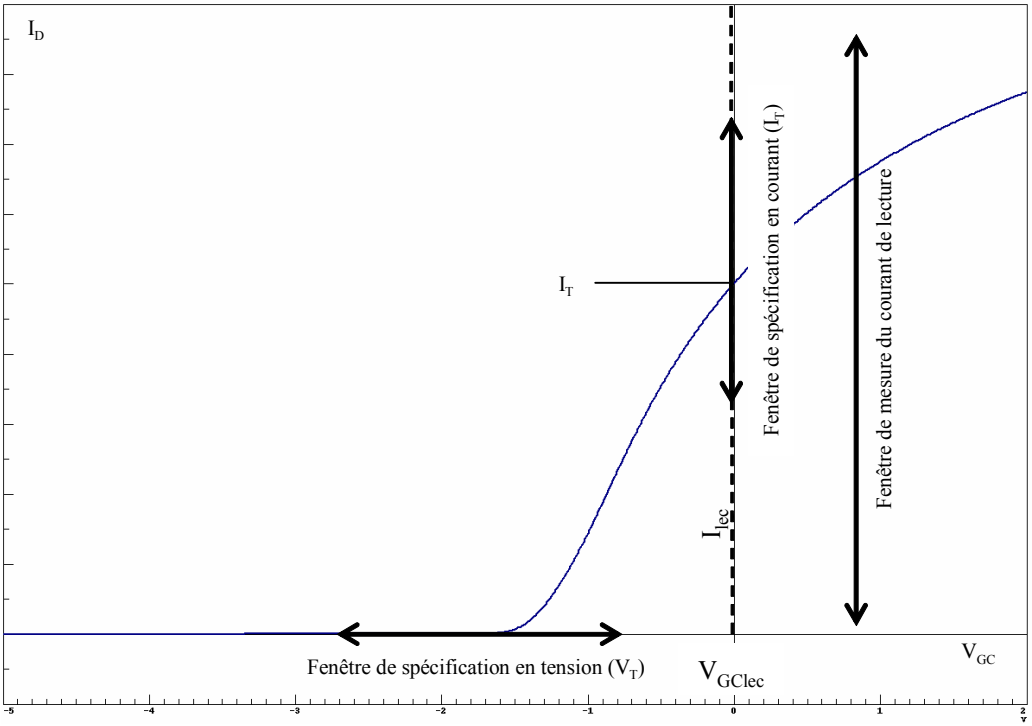


Figure IV. 10 Fenêtre de spécification courant-tension.

2. Structures d'extraction embarquées du courant de seuil

a. Structure d'extraction : génération d'un courant de drain variable

Les conditions d'extraction du courant de seuil sont les suivantes :

- la tension de grille V_{GC} est fixée à une valeur nommée V_{GClec},
- le courant injecté à partir de l'amplificateur augmente par palier de valeur I_{REF}.

Une structure d'extraction composée de sources de courant programmables montées en cascade [UYE97] est utilisée de manière à fournir un courant augmentant par palier. Ces sources de courant élémentaires sont commandées par des interrupteurs de type PMOS (signaux VSW_i). Une sélection progressive de chaque source de courant élémentaire permet d'obtenir un signal évoluant par palier de valeur I_{REF} au niveau du nœud de sortie de la structure I_{read} (figure IV.11).

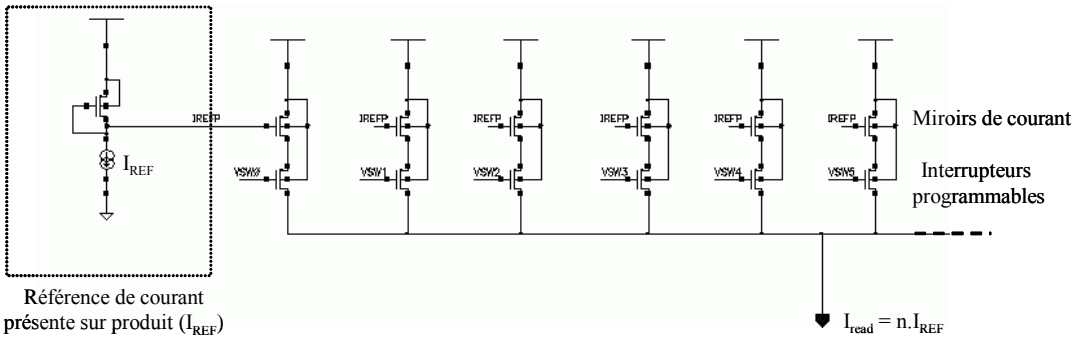


Figure IV. 11 Source de courant programmable par palier.

Là aussi, un appariement précis des références de courant élémentaires (« matching ») est nécessaire de manière à obtenir un pas de variation constant du courant de sortie. De plus, il est à noter que cette structure peut être utilisée pour l'extraction des courants de seuil des trois états électriques du point mémoire EEPROM. Cet autre aspect sera évoqué plus loin.

b. Procédure d'extraction

Les interrupteurs de type PMOS de la figure IV.11 sont contrôlés par un registre à décalage de manière à fournir sur la sortie I_{read} un signal « en escalier » de pas I_{REF}.

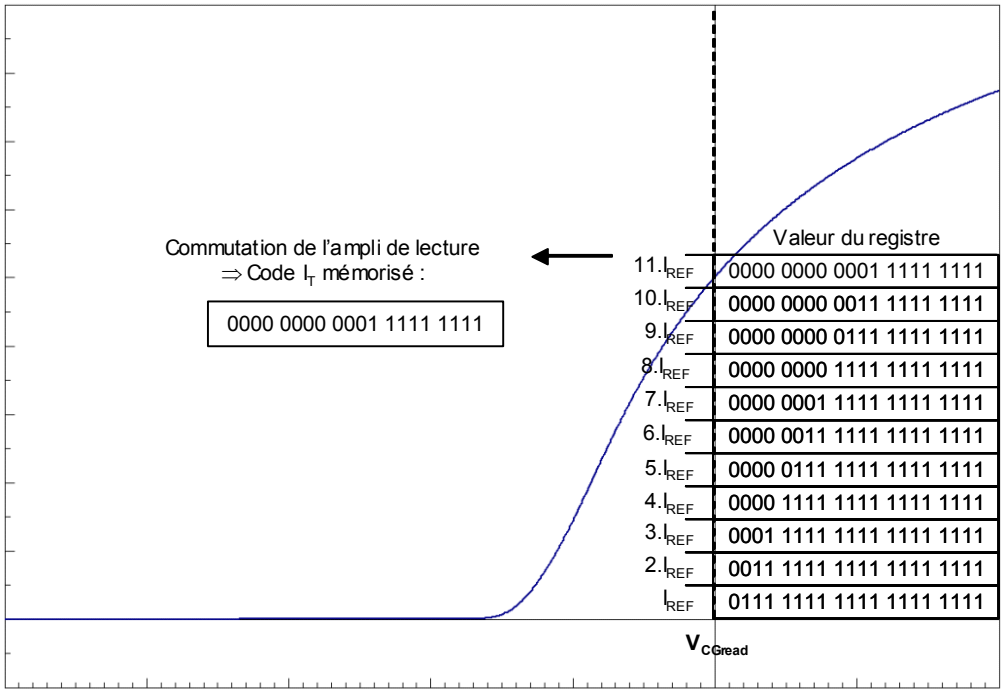


Figure IV. 12 Procédure d'extraction du courant de seuil I_T.

Le registre à décalage utilisé comporte là aussi 20 bits; il est initialisé avec tous ses bits à l'état logique '1', sauf le premier (MSB) qui se trouve à l'état logique '0'.
La figure IV.12 présente un exemple d'extraction du courant de seuil. Le courant de lecture augmente par palier à partir de la valeur I_{REF} en fonction de la valeur du registre à décalage. Le processus de décalage du registre prend fin au moment où l'amplificateur de lecture commute. A ce moment, la valeur du courant de seuil I_T est mémorisée au format numérique et correspond à la valeur binaire contenue dans le registre.

3. Extraction des trois courants de seuil du transistor mémoire

a. Présentation

L'extraction du courant de seuil a été mise en place de manière à obtenir une signature électrique pertinente relative à l'état électrique écrit de la cellule mémoire EEPROM.
Ainsi, il est possible de disposer des valeurs des tensions de seuil dans l'état effacé et vierge de la cellule et du courant de seuil dans l'état écrit de la cellule. La connaissance de ces trois signatures électriques nous permet ainsi de caractériser les trois états de la cellule mémoire EEPROM.
Une deuxième approche consiste à extraire les trois courants de seuil de la cellule mémoire EEPROM : I_{Tecrit}, $I_{Tvierge}$ et $I_{Tefface}$. Dans ce cas, l'extraction des courants de seuil est réalisée pour trois valeurs de tensions de grille différentes comme le montre la figure IV.13.

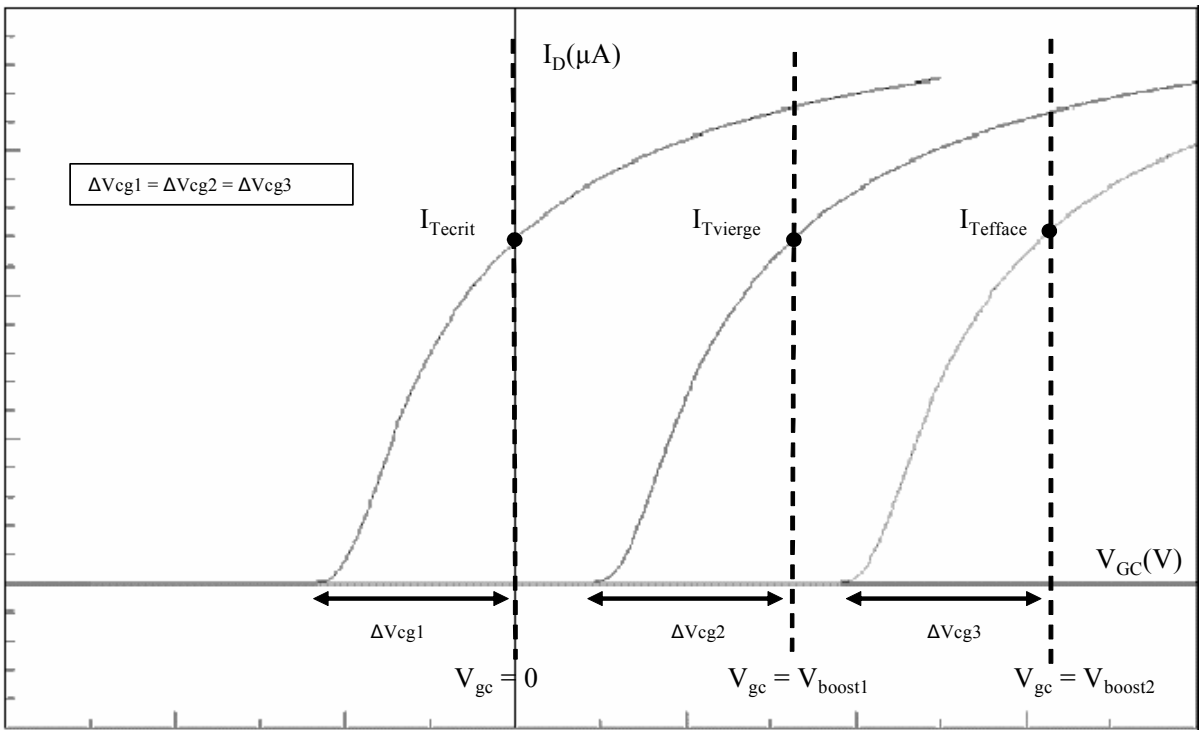


Figure IV. 13 Extraction des trois tensions de seuil du transistor mémoire.

Les trois tensions de grille fixes pour chaque état de la cellule sont :
- $V_{GC} = 0$ dans le cas de l'extraction du courant de seuil écrit,
- $V_{GC} = V_{boost1}$ dans le cas de l'extraction du courant de seuil vierge,
- $V_{GC} = V_{boost2}$ dans le cas de l'extraction du courant de seuil effacé.

b. Utilisation des courants de seuil comme signatures électriques : motivations

Les avantages liés à l'extraction des trois courants de seuil comme signatures électriques de la cellule mémoire EEPROM isolée sur produit sont de deux ordres : d'une part, le surcoût en terme de surface occupée sur la puce sera moins important, puisqu'une seule structure d'extraction est utilisée. D'autre part, comme le montre la figure IV.13, l'extraction des trois courants de seuil est réalisée dans les mêmes conditions ($\Delta V_{cg1} = \Delta V_{cg2} = \Delta V_{cg3}$). Cette dernière caractéristique est très importante puisqu'elle permettra la comparaison de signatures électriques de même dimension, obtenues dans les mêmes conditions de mesure.
Concernant le temps d'accès à la mémoire, ce dernier n'est pas affecté puisque le chemin de lecture n'est en aucun cas modifié par la structure d'extraction qui vient se connecter aux lignes de bit du plan mémoire lors de la mesure.

C. Amélioration des techniques classiques de diagnostic de défauts

1. Le bitmap « analogique »

L'approche classique utilisée pour diagnostiquer les défauts dans le plan mémoire EEPROM consiste à construire une cartographie bit logique du plan mémoire. Comme nous l'avons vu, cette cartographie bit logique du plan mémoire n'est pas suffisante en terme de diagnostic de défauts (cf. §I.A.3b). Cette constatation est particulièrement vraie dans le cas des mémoires EEPROM dont le mode de fonctionnement est purement analogique.
Sachant que les principaux paramètres qui définissent l'état du point mémoire EEPROM sont les tensions ou les courants de seuil (images des tensions de seuil), l'idée principale qui se cache derrière l'extraction de ces valeurs de seuil sur produit est de construire trois cartographies bit analogique.
Chacune de ces cartographies ou « bitmap » fournissant une représentation topologique de toutes les valeurs de seuil du plan mémoire et cela, pour les trois états électriques du plan mémoire.
Les informations obtenues sont :

- une cartographie pour le courant de seuil $I_{Tvierge}$ (ou la tension de seuil $V_{Tvierge}$) des cellules vierges,
- une cartographie pour le courant de seuil $I_{Tefface}$ (ou la tension de seuil $V_{Tefface}$) des cellules effacées,
- une cartographie pour le courant de seuil I_{Tecrit} des cellules écrites.

Pour réaliser une telle cartographie bit analogique, les valeurs numériques des registres à décalage (qui contrôlent les structures d'extraction) sont converties en données analogiques.
Les cartographies bit analogiques peuvent être exploitées de la même manière qu'une cartographie bit numérique, avec des classifications de signatures en fonction des valeurs de seuil I_T (ou V_T).
Les informations obtenues serviront à compléter la représentation « bitmap » logique et à améliorer l'efficacité des étapes de diagnostic de défauts. En effet, ces signatures électriques analogiques sont potentiellement très utiles pour caractériser les cellules défaillantes du plan mémoire.

2. Distribution des valeurs de seuil

Pour compléter le diagnostic de défauts dans les mémoires EEPROM, on peut générer les distributions des courants de seuil (ou des tensions de seuil pour les états électriques vierges et effacés) et cela bien sûr, au sein même du produit EEPROM.
Ces distributions permettront d'évaluer la qualité du processus de fabrication en apportant des informations concernant :

- la robustesse du processus de fabrication,
- les populations de cellules mémoires marginales (dont les valeurs de seuil sont hors spécification),
- la fonctionnalité du plan mémoire.

La suite de ce chapitre est consacrée à la validation des méthodes d'extraction des valeurs de seuil dans les mémoires EEPROM à partir des structures embarquées. Cette étape de validation se fera dans un premier temps par simulation, puis par une série de mesures effectuées sur silicium.
Nous verrons qu'une solution d'extraction embarquée optimisée consistera à extraire uniquement les courants de seuil comme caractéristiques électriques de la cellule mémoire EEPROM.

III. Validation

A. Simulations

1. Architecture d'étude et résultats de simulation

a. Circuit de simulation

Afin de valider l'extraction des valeurs de seuil, des simulations électriques ont été réalisées au moyen de l'outil ELDO [ELD98], sur une mémoire EEPROM réalisée avec les règles de conception de la technologie F8 0,18 µm, produite par la société STMicroelectronics. Cette technologie nécessite une tension d'alimentation V_{dd} de 3,3 V.
Le circuit de simulation est construit autour d'un plan mémoire constitué de 64 cellules EEPROM, plus toute la circuiterie de décodage (cf. Chapitre III, §II.2a). Chaque cellule mémoire étant modélisée par un modèle électrique HDLA. Ce modèle de simulation du transistor mémoire à grille flottante est basé sur un noyau du type MM9. Pour plus d'informations sur ce modèle, on peut se référer à l'article de J.M. Portal [POR02].
Cette matrice élémentaire de cellules EEPROM est composée de mots de 4 bits. Les éléments périphériques pris en compte dans le circuit de simulation (décrit sous forme de « netlist » ELDO) sont :

- le décodeur de ligne et le décodeur de colonne,
- les amplificateurs de lecture relatifs à chacun des quatre bits,
- les bascules de mémorisation (latches),
- les structures d'extraction des valeurs de seuil et leurs circuits associés.

La figure IV.14 schématise le circuit de simulation accompagné des structures d'extraction. Les éléments relatifs aux structures d'extraction apparaissent en grisé sur cette figure.
On constate que le surcoût (en surface silicium) entraîné par l'ajout des structures d'extraction est relativement faible compte tenu de la surface de la puce mémoire. En effet, les structures d'extraction sont principalement composées de :

- une source de courant programmable (utilisant deux transistors PMOS par source de courant élémentaire),
- une source de tension multiple programmable (utilisant quatre transistors MOS par référence de tension élémentaire),
- un registre à décalage configurable (utilisable aussi bien pour le mode d'extraction en courant qu'en tension).

L'intégration de ces structures au sein de l'architecture mémoire nécessite :

- une logique de routage de la tension de grille variable V_{GC},
- la génération de deux signaux V_{boost1} et V_{boost2} au niveau du générateur haute tension,
- une modification de la logique de contrôle de manière à générer les signaux d'extraction des valeurs de seuil dans un mode de test spécifique.

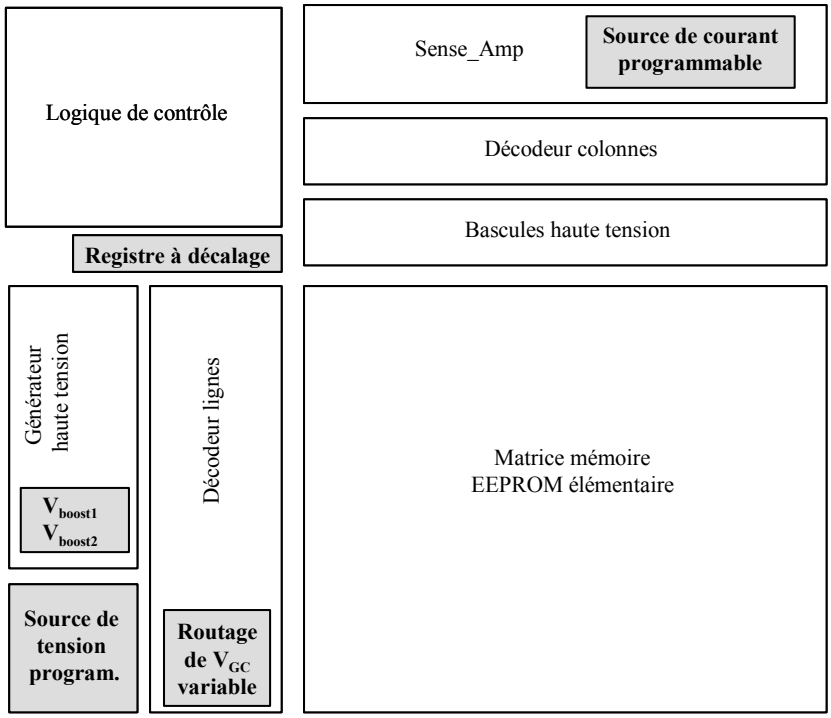


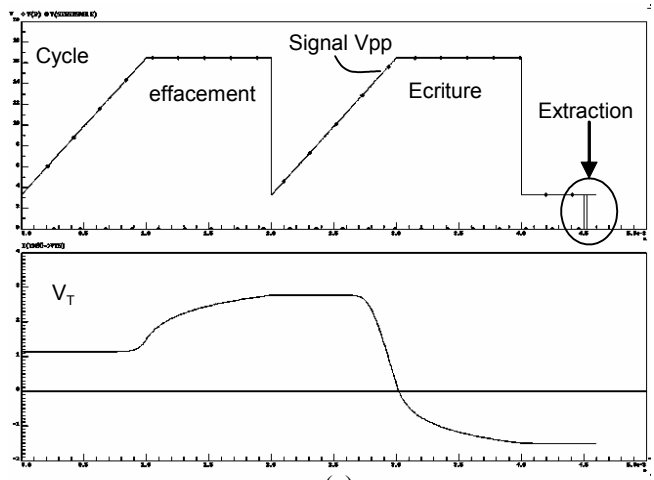
Figure IV. 14 Implémentation des structures d'extraction.

A ce stade, il apparaît clairement que l'extraction des courants de seuil seuls réduirait davantage le surcoût en terme de surface occupée sur la puce mémoire puisque dans ce cas, seule la source de courant programmable, le registre à décalage et la circuiterie de génération de deux signaux V_{boost1} et V_{boost2} seraient à intégrer dans l'architecture.

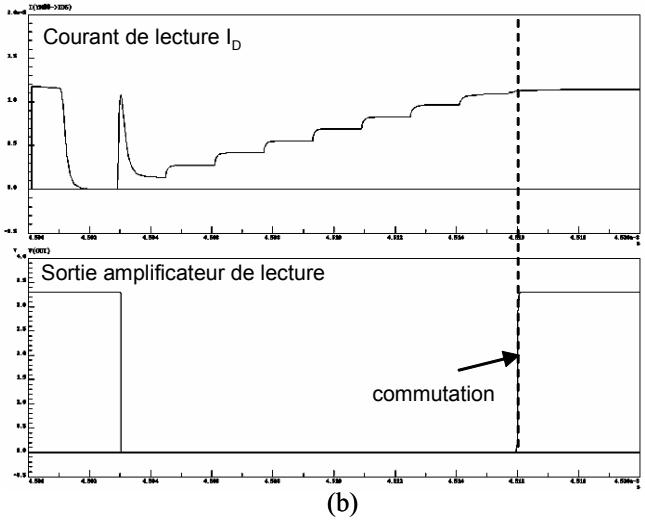
b. Simulations transitoires

Des simulations transitoires ont été réalisées de manière à valider le principe d'extraction des courants de seuil (mais aussi des tensions de seuil). Seuls les résultats relatifs à l'extraction des courants de seuil seront présentés dans cette partie. Une attention particulière sera portée à

la phase de lecture modifiée du point mémoire EEPROM, qui permet l'obtention des valeurs de seuil au format numérique.
La figure IV.15 montre des résultats de simulations transitoires relatives à la phase de lecture modifiée d'une cellule du plan mémoire. La figure IV.15a montre l'évolution de la tension de seuil V_T durant un cycle de programmation (double rampe du signal V_{pp} qui comprend la phase d'effacement et la phase d'écriture). Il apparaît clairement que le temps de lecture imputé à la phase d'extraction du courant de seuil (qui correspond à une lecture modifiée de la cellule mémoire) est relativement faible par rapport au cycle de programmation.
La figure IV.15b est une vue détaillée de la phase d'extraction du courant de seuil de la cellule mémoire. Cette figure fait apparaître une montée par paliers du courant de lecture jusqu'à la commutation de l'amplificateur de lecture (i.e. le courant de lecture à atteint la valeur du courant de seuil de la cellule). A ce moment précis, le transistor mémoire ne laisse plus passer le courant appliqué et la valeur du courant de lecture I_d est mémorisée au format numérique dans le registre à décalage.
Le format numérique des courants de seuil I_T permet de se dispenser de terminal analogique au niveau des moyens de test. Le composant reste ainsi compatible avec des moyens de test entièrement numériques.



(a)



(b)

Figure IV. 15 Simulations transitoires de la phase de lecture modifiée.

Le surcoût en terme de temps de test reste encore très faible puisque le temps de l'opération de lecture modifiée augmente pour occuper 0,5% du temps total de l'opération de programmation. Par ailleurs, le mode de fonctionnement normal de la mémoire est maintenu.
A partir de l'extraction par simulation des courants de seuil, une méthodologie de diagnostic permettant d'exploiter judicieusement les informations extraites d'un ensemble de cellules d'un plan mémoire va être présentée. Comme nous l'avons vu, cette méthodologie de diagnostic de défauts se base sur un « bitmap » analogique et les distributions des valeurs de seuil. Le choix de ces valeurs de seuil étant porté sur les courants de seuil, notre méthodologie de diagnostic se basera uniquement sur l'extraction des courants de seuil.

2. Bitmap analogique en courant

La figure IV.16 représente les cartographies bit analogiques obtenues pour la simulation du plan mémoire 64 bits constitué de mots de 4 bits. Les figures IV.16a, IV.16b et IV.16c représentent respectivement les cartographies analogiques relatives au plan de cellules mémoires vierges, au plan de cellules écrites et au plan de cellules effacées. Les valeurs des courants de seuil sont exprimées en « µA » pour les trois états du plan mémoire.
Dans ce plan mémoire, la moitié des cellules a été conçue avec la géométrie cible de la technologie, alors que l'autre moitié a été dimensionnée de manière aléatoire avec des paramètres géométriques qui présentent une variation comprise entre 5% et 40% par rapport à la valeur cible. Ce dimensionnement différent des cellules du plan mémoire a été réalisé de manière à mettre en évidence l'impact de la variation de la géométrie des cellules mémoires sur les courants de seuil.

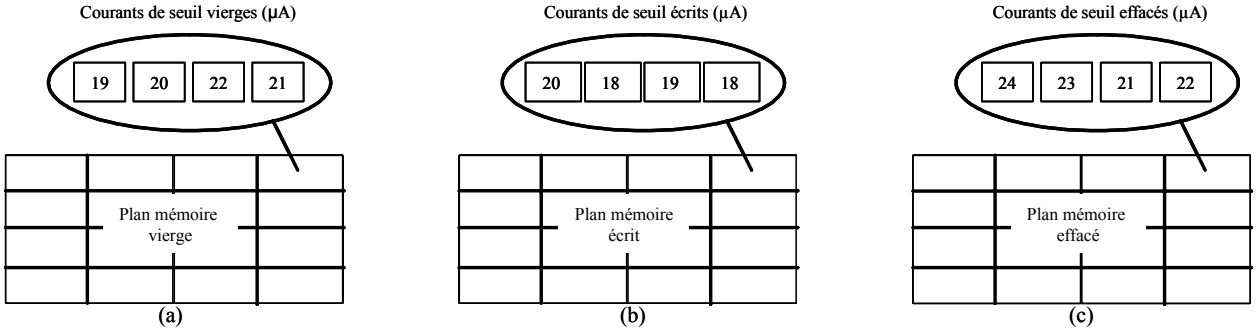


Figure IV. 16 Cartographies analogiques en courant pour les trois états du plan mémoire.

Dans le présent exemple, une cartographie bit logique ou « bitmap » logique ne fait apparaître aucun bit défaillant. Cependant, la cartographie bit analogique révèle l'existence de cellules pour lesquelles les courants de seuil sont hors spécification (ou à la limite des spécifications), et permet en outre de localiser ces cellules. A partir de ces valeurs analogiques hors spécification, il est envisageable de mettre en œuvre des outils de diagnostic spécifiques pouvant servir à identifier les paramètres géométriques défectueux.

3. Distribution des courants de seuil

Il est également possible de générer les distributions des courants de seuil pour les trois états électriques du plan mémoire, de manière à évaluer la dispersion des courants de seuil et comparer les distributions obtenues avec les spécifications.

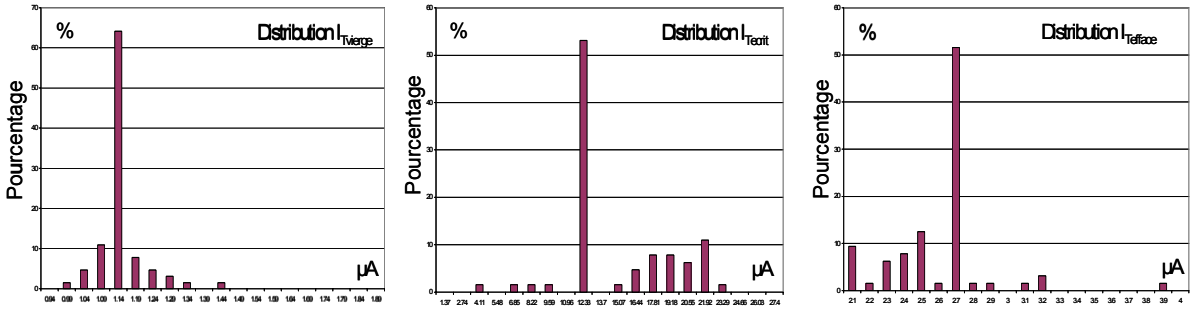


Figure IV. 17 Distribution des courants de seuil pour les trois états du plan mémoire.

Les figures IV.17a, IV.17b et IV.17c constituent des exemples de distributions des courants de seuil obtenus pour le plan mémoire simulé. Pour les trois états du plan mémoire, les distributions obtenues mettent en évidence une raie centrale relative aux cellules conçues avec la géométrie cible de la technologie considérée et des cellules marginales. Ces cellules marginales, dont les géométries différent aléatoirement des valeurs cibles, forment des queues de distribution (cas des cellules vierges) ou bien mettent en évidence des populations entières de cellules marginales (doubles distributions pour le plan mémoire écrit et effacé).
Ces distributions en courant de seuil donnent des informations relatives à la répartition des défauts sur l'ensemble du plan mémoire. Plus encore, ces distributions reflètent la dispersion des paramètres de fabrication technologiques et fournissent par conséquent de précieuses informations aux ingénieurs « process ». Ce qui permet un contrôle régulier du processus de fabrication.

B. Validation silicium

1. Le véhicule de test EEPROM 0,18 µm

a. Présentation

Il existe des structures de test embarquées ou « Test chip » qui sont des véhicules de test conçus pour l'introduction de nouvelles technologies. La figure IV.18 représente un schéma bloc fonctionnel du véhicule de test EEPROM qui accompagne le processus de fabrication de la technologie EEPROM 0,18 µm (mémoire EEPROM embarquée pour les technologies cartes à puces). Le circuit est composé d'une matrice de cellules 512 Kbits avec uniquement la circuiterie de décodage et de lecture. Tous les signaux de programmation sont fournis extérieurement, ce qui procure une flexibilité maximale en terme de caractérisation. Ainsi, une caractérisation précise d'une cellule du plan mémoire est possible.
L'accès aux valeurs de seuil des cellules du plan mémoire par exemple, est possible via un accès direct aux colonnes (ou « bit lines ») de la mémoire. Pour cela, le véhicule de test est d'abord placé dans un mode de test spécifique où la plupart des circuits périphériques sont court-circuités.
On constate à partir de la figure IV.18 que les signaux de programmation (V_{PP} et V_{boost}) et de lecture (I_{REF_10uA}, I_{REF_CDMA} et V_{REF}), qui sont des signaux internes dans le cas des produits EEPROM, sont accessibles directement à partir des broches externes du composant.
Pour des raisons de confidentialité, nous n'irons pas plus loin dans la description de ce véhicule de test.

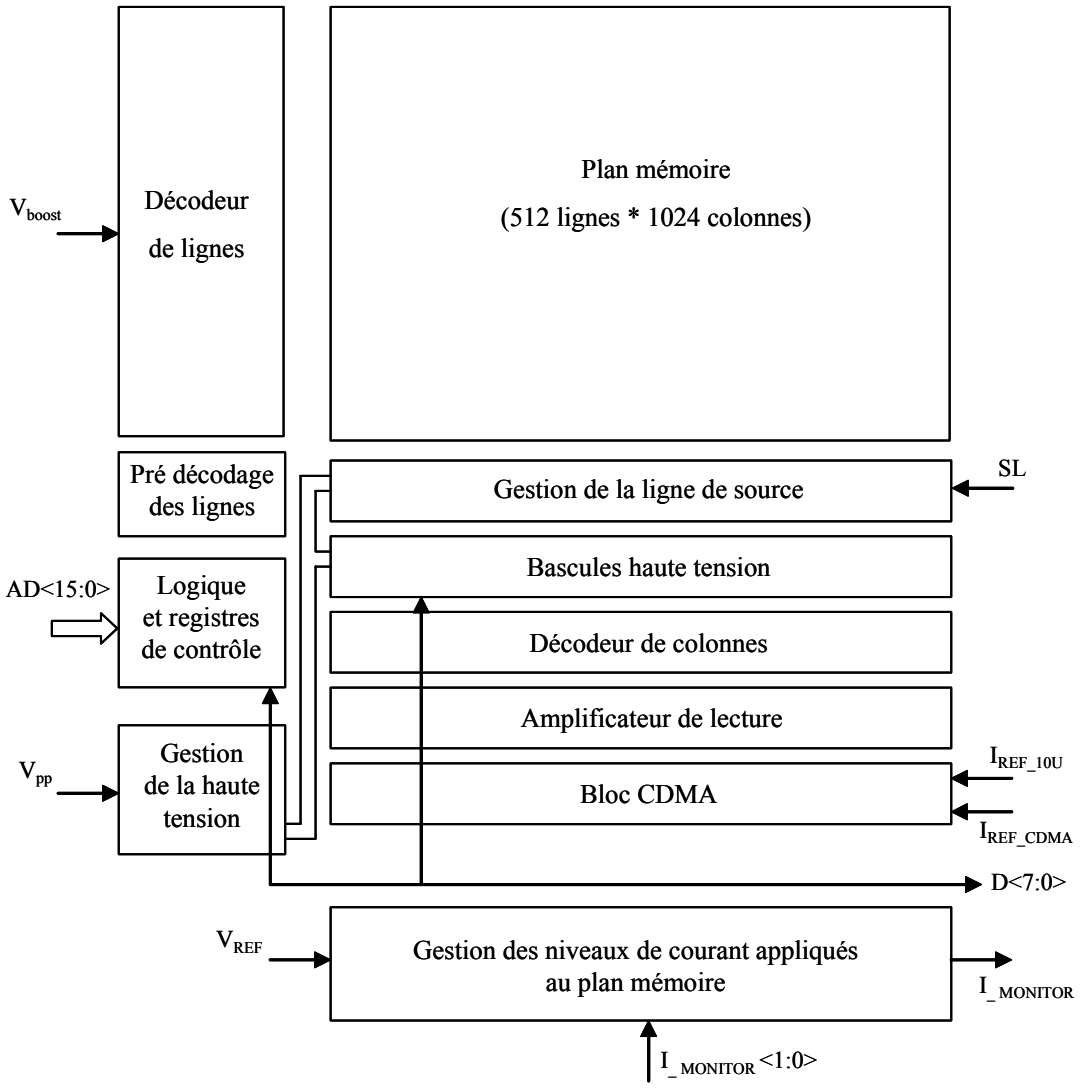


Figure IV. 18 Le véhicule de test EEPROM 0.18 µm.

b. Caractéristiques du véhicule de test

Le véhicule de test EEPROM permet un accès aux nœuds internes d'une cellule du plan mémoire, ce qui permet d'extraire les caractéristiques $I_d(V_{GC})$ d'une cellule isolée par exemple, et cela pour ses trois états électriques comme le montre la figure IV.19.
Dans le cadre de notre étude, l'utilisation de ce véhicule de test va nous permettre de valider, dans un premier temps, notre technique d'extraction des courants de seuil.
En effet, le signal appliqué sur la grille de contrôle V_{GC} ainsi que le signal de drain de chaque cellule du plan mémoire peuvent être fournis à partir de broches externes. Ce qui va nous permettre d'augmenter par paliers le courant de lecture d'une cellule adressée pour une tension de grille V_{GC} préalablement fixée.
La commutation de l'amplificateur de lecture (signaux D<7 :0>) signifiera que le courant de lecture aura atteint le courant de seuil de la cellule adressée.

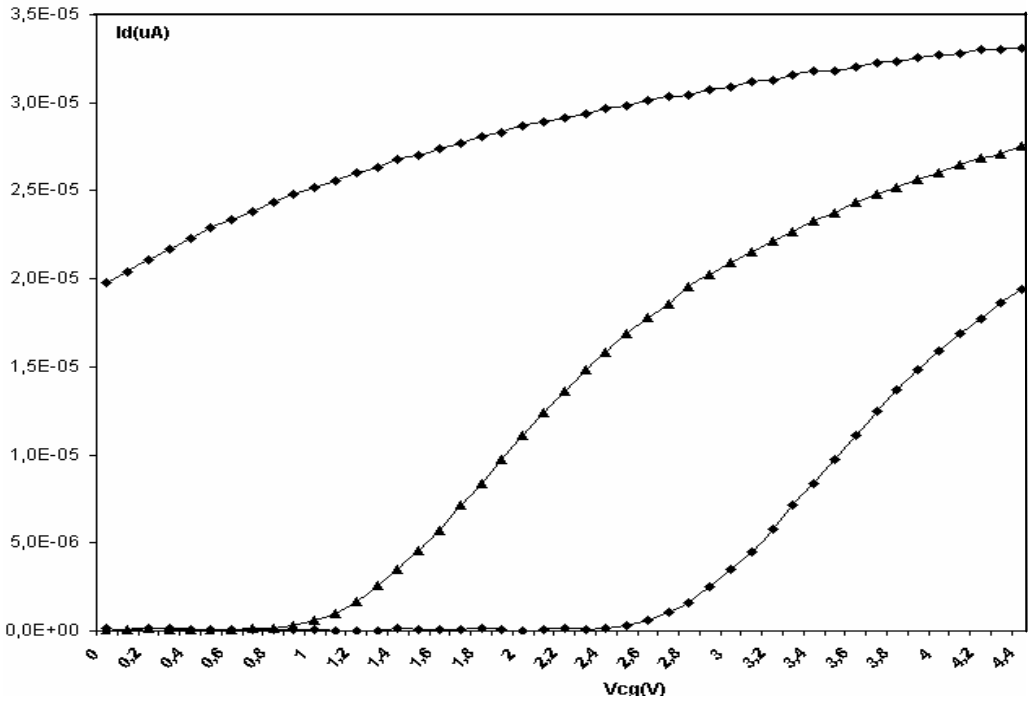


Figure IV. 19 Extraction des caractéristiques Id(V_{GC}) pour les trois états de la cellule EEPROM 0.18µm.

2. Algorithme d'extraction du courant de seuil

La validation silicium de notre technique d'extraction passe par le développement d'une routine de test qui correspond à une lecture modifiée du point mémoire EEPROM. Cette routine d'extraction des courants de seuil sera ensuite implémentée sur un testeur VLSI.
L'algorithme d'extraction du courant de seuil est décrit figure IV.20. Le processus d'extraction commence par la définition de la partie mémoire sur laquelle les mesures seront effectuées (ligne initiale « $Ligne_i$ » et ligne finale « $Ligne_f$ », colonne initiale « Col_i » et colonne finale « Col_f »). Ensuite, tous les paramètres de lecture sont fixés, comme la tension de grille V_{GClec}, la fenêtre de mesure en courant limitée par le paramètre I_{fin} et le pas de variation du courant de lecture I_{REF}.
Pour chaque pas de variation du courant de lecture I_{lec} et pour chaque adresse, le nombre de cellules du plan mémoire qui conduisent le courant appliqué (i.e. cellules dont le courant est inférieur à la valeur du courant de seuil) est mémorisé dans une variable nommée « Nbr_Cellules ».
Lorsque toute la partie du plan mémoire étudiée a été analysée, il est possible de construire les distributions des courants de seuil. Ces distributions sont en général extraites en prenant en compte la totalité du plan mémoire de manière à obtenir des informations statistiques.
Dans le cas du « bitmap » analogique, l'adresse de la cellule dont le courant de seuil est atteint est mémorisée de manière à construire une cartographie bit analogique des courants de seuil. Le « bitmap » analogique est généralement réalisé sur une petite partie du plan mémoire pour des raisons liées aux capacités de stockage d'informations sur le testeur.
Ainsi, le « bitmap » analogique est utilisé de manière à analyser d'éventuelles régions de la puce ou les valeurs des courants de seuil seraient anormales.

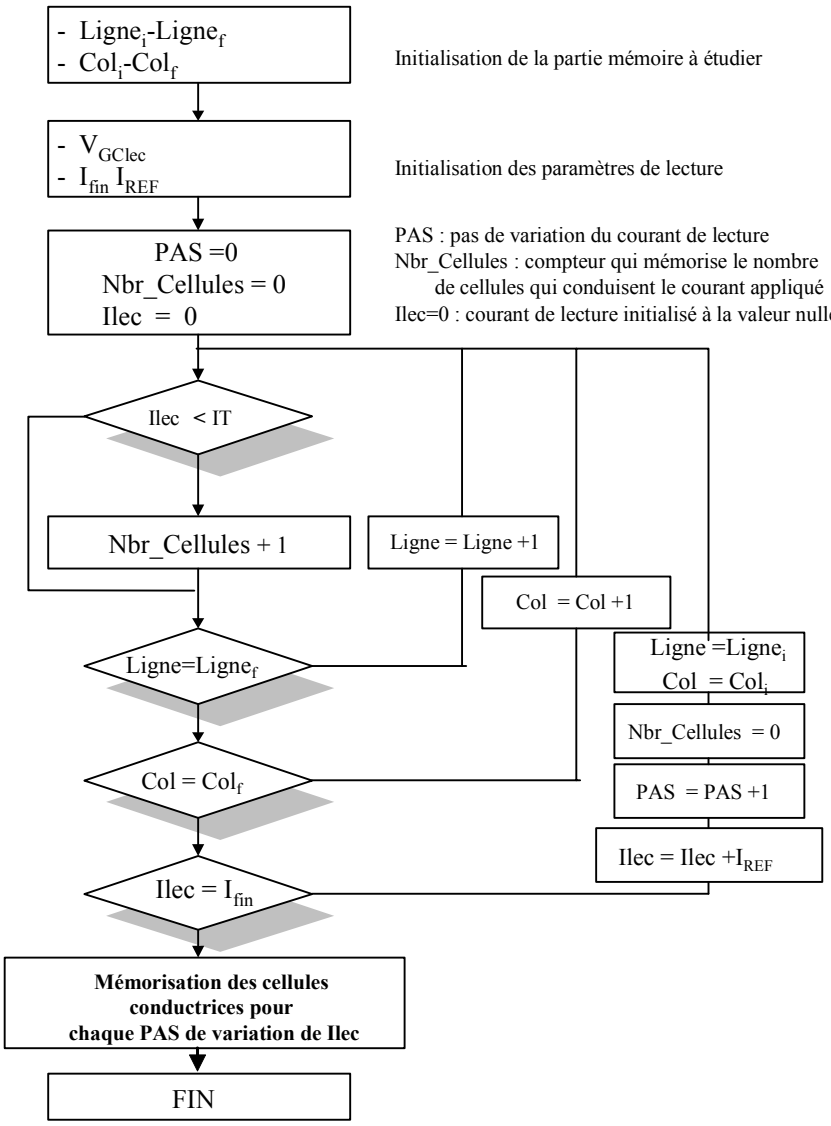


Figure IV. 20 Algorithme d'extraction des courants de seuil.

3. Bitmap « analogique » en courant

L'implémentation informatique de l'algorithme d'extraction des courants de seuil sur un testeur de circuit VLSI (de type « QT-200 »), spécifique aux mémoires non volatiles, a permis de valider dans un premier temps la notion de « bitmap » analogique en courant.
Les figures IV.21a, IV.21b et IV.21c sont une extraction des courants de seuil, exprimés en µA pour les trois états du plan mémoire. Les mesures ont été effectuées sur une partie du plan mémoire du véhicule de test EEPROM 0,18 µm. Cette partie du plan mémoire est constituée de deux colonnes de mots de 8 bits et de quatre lignes.
On distingue clairement des variations (de l'ordre du µA) des valeurs des courants de seuil pour une représentation donnée.
Cette représentation sous forme analogique des cellules du plan mémoire permet bien sûr de diagnostiquer plus efficacement les défaillances affectant le plan mémoire. En effet, contrairement au « bitmap » dit logique où la seule information obtenue pour une cellule défaillante (ou un groupe de cellules) est sa localisation, ici, une cellule défaillante peut être caractérisée par trois valeurs analogiques. Ces trois dernières valeurs évoluant dans une fenêtre de variation représentative de la qualité du processus de fabrication.

A partir de ces valeurs spécifiques, des outils de diagnostic peuvent être mis en place de manière à déterminer l'origine d'une défaillance induite par la variation d'un paramètre géométrique par exemple (cf. Chapitre II).

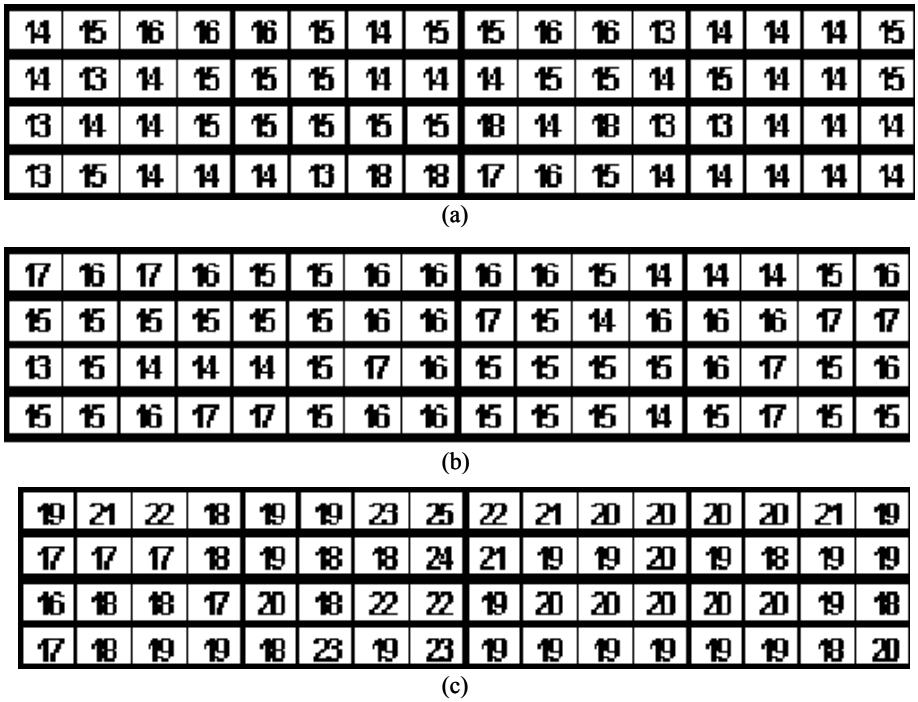


Figure IV. 21 « Bitmap » analogique du plan mémoire vierge (a), effacé (b) et écrit (c).

Ces valeurs analogiques évoluent durant la durée de vie du produit EEPROM. Cette évolution est mise en évidence lors des tests en endurance sur des cellules isolées, situées dans les lignes de découpe. Je rappelle que le test d'endurance consiste à effectuer plusieurs cycles d'effacement/écriture sur une cellule élémentaire et à prélever la tension de seuil correspondante à chaque nombre de cycles (en écriture et en effacement). La différence entre la tension de seuil en écriture et celle en effacement est appelée fenêtre de programmation.
Généralement, on constate une évolution des tensions de seuil (et donc des courants de seuil) qui entraîne une fermeture de la fenêtre de programmation (figure IV.22).

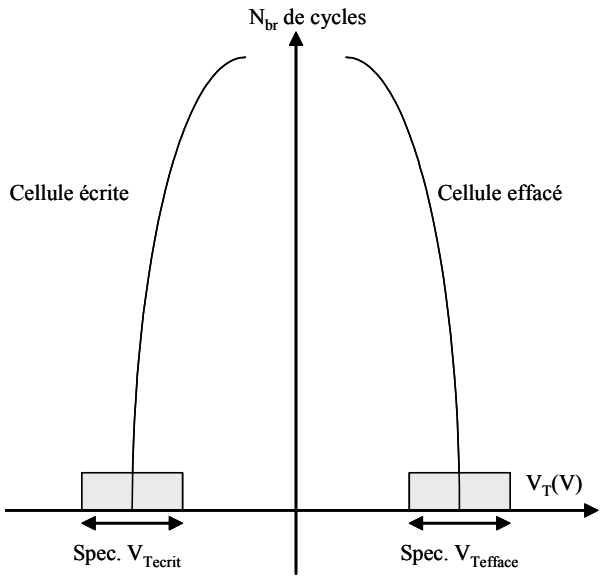


Figure IV. 22 Fermeture de la fenêtre de programmation.

Grâce au bitmap analogique et en résonnant cette fois en courant, la position et le nombre des cellules mémoires dont les courants de seuil sont hors spécification ou à la limite des spécifications peuvent être connues avant le test en endurance.
Ces cellules auront une probabilité plus élevée d'engendrer une défaillance lors du fonctionnement de l'EEPROM. Ces informations analogiques sont donc très importantes en terme de fiabilité de la mémoire.

4. Distribution des courants de seuil

Toujours à partir de l'algorithme décrit à la figure IV.20, les distributions en courants pour les trois états du plan mémoire ont été mesurées sur le véhicule de test EEPROM 0,18 µm.
Les figures IV.23a, IV.23b et IV.23c représentent le résultat de l'extraction des distributions de courants de seuil en écriture, en effacement et pour le plan mémoire vierge. Ces distributions ont été mesurées sur la totalité du véhicule de test.
Pour chacune distribution, le nombre de cellules dont le courant de seuil est atteint apparaît en ordonnée et le courant appliqué est donné en abscisse.

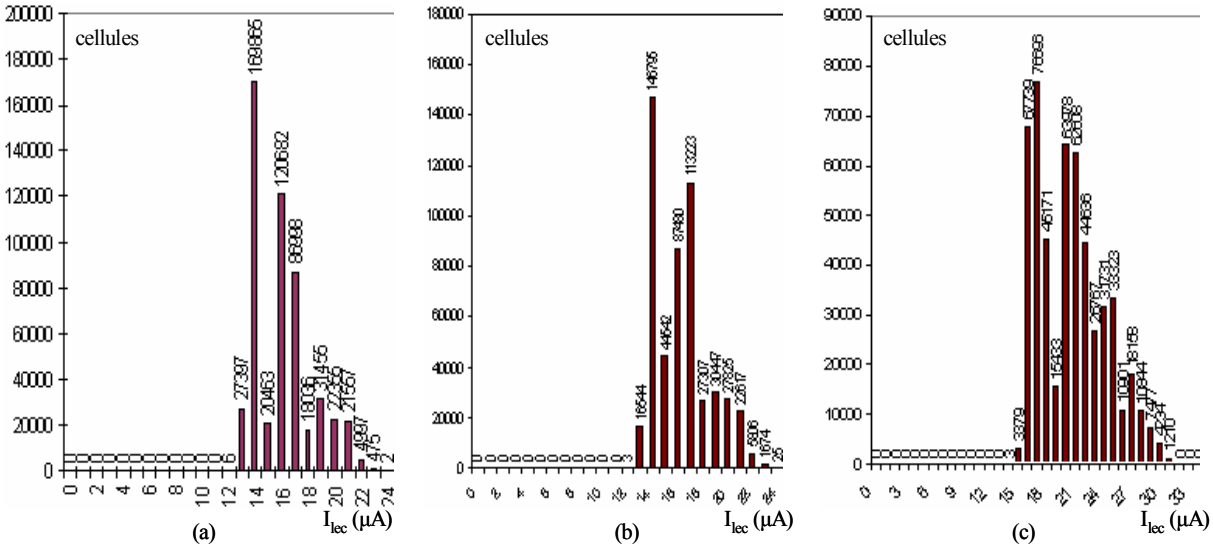


Figure IV. 23 Distributions en courants de seuil écrites(c), effacées (b) et vierges (a).

Ces distributions sont censées refléter la qualité du processus de fabrication (étalement des distributions) ou mettre en évidence la dérive d'un paramètre technologique. De manière à valider cette hypothèse l'impact de la variation d'un paramètre de fabrication spécifique de la cellule EEPROM sur les distributions des courants de seuil est présenté dans la partie suivante.

5. Impact d'un paramètre du processus de fabrication sur les courants de seuil

De manière à montrer l'impact d'un paramètre du processus de fabrication des mémoires EEPROM sur les courants de seuil, des mesures ont été effectuées sur des tranches de silicium présentant des épaisseurs d'oxyde tunnel T_{tun} différentes.
Dans les mémoires EEPROM, le transfert de charges entre le drain et la grille flottante se fait par injection courant Fowler-Nordheim à travers l'oxyde tunnel. La fiabilité des mémoires EEPROM dépend donc de manière importante de l'épaisseur et de la qualité de cet oxyde.
L'intégrité de ces fines couches d'oxyde tunnel est généralement contrôlée lors de tests Q_{bd} effectués sur de larges capacités [BAG87]. Cependant ces tests sont indicatifs de la qualité

intrinsèque de l'oxyde tunnel et ne donnent pas d'informations relatives au comportement d'une large population de cellules mémoires qu'on retrouve embarqués sur produit.
La figure IV.24 montre l'impact de l'épaisseur d'oxyde tunnel sur les distributions en courants. Ces mesures ont été effectuées sur quatre tranches de silicium différentes présentant des épaisseurs d'oxyde tunnel variant de 56 Å à 70 Å. Les distributions en courant ont été relevées pour le plan mémoire EEPROM dans un état électrique effacé.

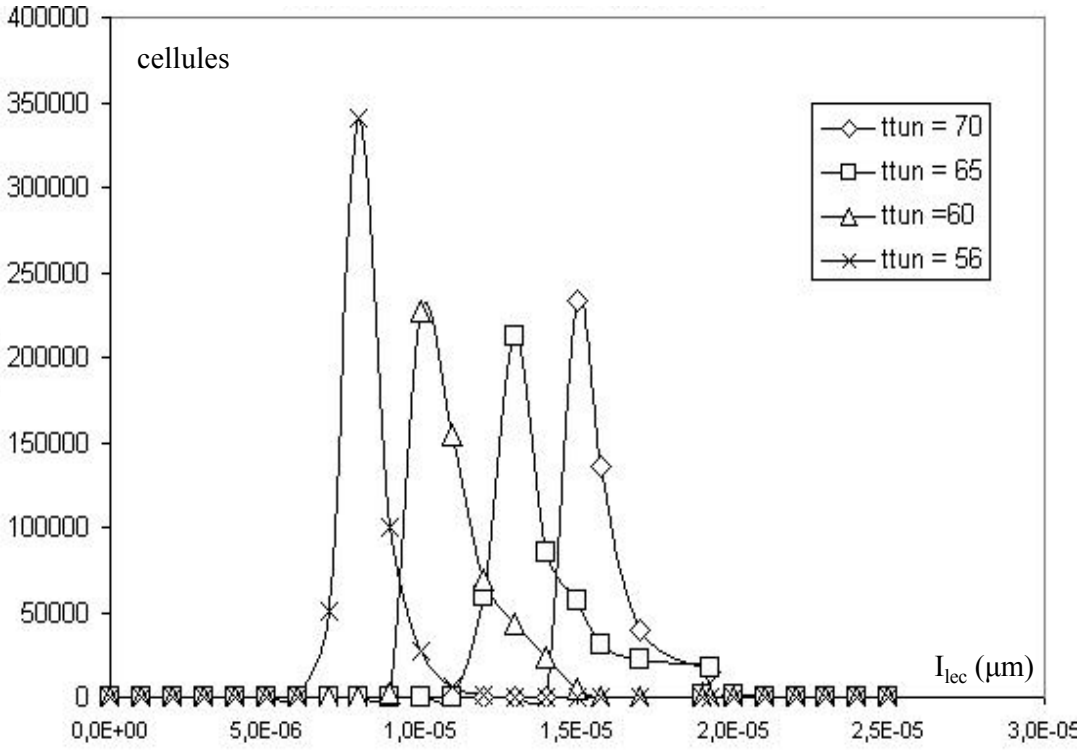


Figure IV. 24 Impact de l'épaisseur de l'oxyde tunnel sur les distributions en courant.

On constate un décalage des distributions en courants de seuil pour chaque valeur de l'épaisseur de l'oxyde tunnel. L'impact de ce paramètre sur les distributions en courant de seuil est donc clairement mis en évidence à partir de cette série de mesure.

C. Conclusion

Les mémoires non-volatiles de type EEPROM sont constituées des cellules dont le principe de fonctionnement est purement analogique. Ceci conduit à des problèmes critiques au niveau du contrôle du procédé de fabrication mais aussi lors de la phase de conception.
Des structures d'extraction intégrées au composant permettent d'extraire des signatures analogiques pertinentes représentatives de chaque cellule du plan mémoire (courants de seuil dans le cadre de notre étude). Les courants de seuil sont extraits durant le flot de test fonctionnel avec un faible surcoût en termes de temps de test.
Il a aussi été montré que la surface occupée sur la puce par les structures d'extraction était très faible. De plus, les courants de seuil sont obtenus sous forme numérisée, ce qui permet une compatibilité totale avec des équipements de test entièrement numériques.
Il est ainsi possible à partir de ces nouvelles données de développer une méthodologie de diagnostic basée sur une cartographie bit analogique, pouvant être utilisée de manière à compléter une cartographie bit numérique classique. Il en résulte une analyse plus efficace du comportement analogique de chaque cellule du plan mémoire.

Cette méthodologie de diagnostic permet également d'obtenir les distributions des courants de seuil sur l'ensemble du plan mémoire, ce qui peut être utile pour la qualification d'un procédé de fabrication.
Cette approche a été entièrement validée sur silicium à partir du véhicule de test EEPROM 0,18 µm.

Conclusion Générale

Ce manuscrit fait état d'une étude approfondie de différentes méthodes de test et de diagnostic de défauts spécifiques aux mémoires non volatiles de type EEPROM. D'un point de vue général, notre approche du test et du diagnostic de défauts s'est basée sur l'obtention de signatures électriques de type analogique comme les courants ou les tensions de seuil. Ces signatures, représentatives de l'état de chaque cellule du plan mémoire, sont à la base d'une méthodologie de diagnostic de défauts qui cible à la fois la cellule EEPROM isolée et le plan mémoire EEPROM.
La technique de diagnostic de défauts géométriques dans la cellule EEPROM a fait l'objet d'un brevet STMicroelectronic : "*Procédé de modélisation mathématique de composants électroniques et son utilisation pour la simulation, la détermination de géométries et le diagnostic de défauts dans le composant* ", *N° de dépôt 0208652, N° de dossier 02-RO-029, 2002* (*Portal J.M., Forli L., Aziza H., Née D).*

Ce procédé nous a permis de générer un modèle mathématique avec lequel nous avons pu caractériser les dépendances de la tension de seuil vis à vis des paramètres géométriques de la cellule mémoire. Ce modèle mathématique est basé sur un modèle électrique compact de transistor à grille flottante. Un plan de simulation (DOS) a été utilisé pour générer le modèle mathématique. Cette approche permet d'étudier de manière extrêmement rapide, à des fins de diagnostic, toute influence d'un paramètre d'entrée quelconque de la cellule sur une grandeur électrique de sortie (tension de seuil dans le cadre de notre exemple d'application). Une validation silicium de notre méthodologie a montré l'efficacité de notre modèle mathématique. Ceci nous a poussé à développer un outil logiciel basé sur une interface graphique de manière à automatiser notre méthodologie de diagnostic de défauts dans les EEPROM.
Cependant, ce modèle reste ouvert puisque d'autres paramètres peuvent être pris en compte au niveau du modèle mathématique de la tension de seuil (tension de programmation V_{pp}, doses d'implantation, autre paramètres géométriques…) de manière à diagnostiquer d'une manière générale une variation d'un ou de plusieurs des paramètres d'entrée susceptibles d'entraîner une variation hors spécification d'une ou de plusieurs des grandeurs électriques de sortie.
De plus, cette technique peut être utilisée de manière à cibler d'autres composants mémoire tels que les mémoires FLASH et les mémoires eDRAM. Les applications potentielles de ce procédé de modélisation sont donc multiples et peuvent intéresser particulièrement le domaine des circuits analogiques.

L'analyse des défauts au niveau de la matrice mémoire a débuté par la définition d'une architecture d'étude composée d'un plan mémoire élémentaire et de tous ses circuits périphériques. Ensuite, à partir de la connaissance du processus de fabrication des mémoires EEPROM et de l'étude du dessin des masques de la technologie considérée, l'impact des différents défauts pouvant potentiellement affecter le fonctionnement du plan mémoire a été mis en évidence : le comportement électrique de mémoires complètes ou de parties de mémoires incluant divers types de défauts a été observé à partir de l'extraction des tensions de seuil de toutes les cellules du plan mémoire après simulation.

Cette étude a été motivée par les raisons suivantes : d'une part, dans le cas des mémoires EEPROM embarquées (cartes à puces), la majorité des défauts affecte le plan mémoire EEPROM. Cela s'explique par la complexité du mécanisme de fonctionnement de la cellule élémentaire (transfert de charge de type FN à travers un oxyde fin) qui rend les performances associées à la partie EEPROM très sensibles au processus de fabrication. D'ailleurs, ces observations ont été confirmées par les nombreux résultats de test obtenus sur ce type de produit. D'autre part, les mémoires constituent un véhicule test puissant. En effet, un défaut dans la mémoire tend à causer une défaillance très proche de sa localisation physique et cela, grâce aux caractéristiques de symétrie offertes par le plan mémoire.

Nous avons vu que les mémoires EEPROM présentent un mécanisme de mémorisation de type analogique faisant appel à un basculement des tensions de seuil, mais sont testées en utilisant des algorithmes classiques de test comme ceux appliqués aux mémoires vives (RAM), et sont diagnostiquées au moyen d'une cartographie bit logique.
Dans le but d'améliorer les étapes de diagnostic, une méthodologie de test basée sur des moyens intégrés d'extraction a été mise en place dans le but d'acquérir des signatures électriques significatives (tensions et courants de seuil). Dans le cas de l'extraction des courants de seuil, deux nouvelles informations très utiles pour caractériser un procédé de fabrication (effets des changements de chimie, des changements d'équipement…) sont désormais disponibles : la première information est une cartographie bit analogique qui comprend trois représentations :

- une pour le courant de seuil $I_{Tvierge}$ d'une cellule vierge,
- une pour le courant de seuil $I_{Tefface}$ d'une cellule effacée,
- une pour le courant seuil I_{Tecrit} d'une cellule écrite.

Ces cartographies bit analogiques peuvent être exploitées de la même manière qu'une cartographie bit numérique (« bitmap »), avec des classifications de signatures en fonction des valeurs de seuil.
Cette nouvelle méthodologie de diagnostic permet également d'obtenir les distributions des courants de seuil pour les trois états électriques de l'ensemble du plan mémoire, ce qui peut être utile pour la qualification d'un procédé de fabrication. Toutes ces informations supplémentaires, qui vérifient l'état analogique de chaque cellule du plan mémoire, sont obtenues sous un format numérique lors de flots de tests fonctionnels standard modifiés.
La mise en place des structures d'extraction de signatures électriques a fait l'objet d'un brevet STMicroelectronic : *"Circuit de caractérisation des tensions de seuils dans les mémoires non volatiles sous format numérique et traitement en vue du diagnostic des données obtenues", N° de dépôt 0306184, N° de dossier 03-RO-059, 2003* (*Portal J.M., Aziza H., Née D)*. Aux vues des résultats obtenus sur silicium à partir d'un véhicule de test EEPROM 0,18 µm (« bitmap analogique » et distribution des courants de seuil), il est prévu la mise en place des structures d'extraction directement sur les produits cartes à puce embarquant de la mémoire EEPROM.

Les perspectives envisageables à moyen terme se basent sur l'acquisition d'un testeur de circuit VLSI Agilent 93000 par le Laboratoire de Matériaux et de Microélectronique de Provence (L2MP). Cette acquisition va nous permettre de développer des programmes de test « intelligents » qui seront capables d'extraire tous les paramètres analogiques représentatifs des cellules défaillantes d'un plan mémoire, en plus des résultats de test classiques. Cela passera par une modification des programmes de test standard de manière à prendre en compte le mécanisme de fonctionnement spécifique des mémoires non volatiles
La finalité est de diagnostiquer le plus rapidement possible tout dysfonctionnement du circuit mémoire affectant le rendement de fabrication.

Pour conclure, je dirai que ce travail de thèse a fait appel à une multitude de compétences et domaines de la microélectronique (test de circuits, technologie de fabrication, analyse de défaillance, conception de circuits…), ce qui l'a rendu d'autant plus enrichissant et passionnant.

Bibliographie

[APP70] J. A. APPELS, E. KOOL, M. M. PAFFEN, J. J. H. SCHATORJE, W. H. C. G. VERKUYLEN, "Local Oxidation of Silicon and its application in Semiconductor-Device", Philips Research Reports, Vol 25, pp.119-132, 1970.

[ARO99] N. ARORA, "MOSFET Models for VLSI circuit simulation. Theory and practice", New York, Springer-Verlag, ISBN 3-211-82395-6, 1993.

[BAG87] D.A. BAGLEE, T. SUGAWARA, S. FUKAWA, K. MORI, L.M. BELLAY, T. MILLER, "The Effects of Processing on EEPROM Reliability", Proc. IEEE International Reliability Physics Symposium, pp.93-96, 1987.

[BAR76] W. BARRACLOUGH, A. CHIANG, W. SOHL, "Techniques for testing the microcomputer family", Proceedings of the IEEE, Vol 64, no 6, 1976.

[BAR98] S. BARBERAN, F. DUVIVIER, "Management of Critical Areas and Defectivity Data for Yield Trend Modeling", Proc. IEEE Int. Symp. On Defect and Fault Tolerance in VLSI Systems, pp.17-25, 1998.

[BCH99] M. BUCHER *et al*, "The EPFL-EKV Mosfet Model, version 2.6", EPFL, Tech. Rep.1999, Internet source : http://legwww.epfl.ch/ekv/.

[BER68] E. R. BERLEKAMP, "Algebraic Coding Theory", New York: McGraw-Hill, 1968.

[BOU99] R. BOUCHAKOUR, "Physique des composants", Cours : Université Aix-Marseille I, Laboratoire Matériaux et Microélectronique de Provence (L2MP), 1996.

[BRE79] M. BREUER, A. FRIEDMAN, "TEST/80: A proposal for an advanced automatic test generation system", IEEE Autotestcon., 1979.

[BRE80] M. BREUER, A. FRIEDMAN, "Functional level primitives in test generation", IEEE Transactions on Computers, Vol C-29, 1980.

[BRO98] W .D. BROWN, J. E. BREWER, "Non Volatile semiconducteur memory technologie", IEEE PRESS, New York, 1998.

[CAN00] P. CANET, R. BOUCHAKOUR, N. HARABECH, P. BOIVIN, J. M. MIRABEL, C. PLOSSU, "Study of signal programming to improve EEPROM cell reliability", IEEE Cir. & Sys., Vol 3, pp. 1144-1147, 2000.

[CAP99] P. CAPPELLETTI, C. GOLLA, P. OLIVO, E. ZANONI, "Flash memories", Kluwer Academic Publishers, 1999.

[CAR89] G. le CARVAL, "Contribution de l'étude physique de l'injection par porteurs chauds à travers l'oxyde de grille des TMOS dans le cadre du fonctionnement des mémoires non-volatiles", Thèse, Institut National Polytechnique de Grenoble, 1989.

[CAR93] E. CARMAN, C. LAWRENCE, R. NAIR, G. SANCHEZ, "Isolating the Killer Defect: Process Analysis using Particle Map to Probe Map Correlation", Advanced Semiconductor Manufacturing Conf. and Work., ASMC 93 Proceedings, IEEE/SEMI, pp.198-200, 1993.

[CHA87] C. C. CHAO, M. H. WHITE, "Characterization of charge injection and trapping in scaled SONOS/MNOS memory devices", Sol. St. Elect., Vol 30, p.307, 1987.

[CHE69] A. CHEN, "Redundancy in LSI memory array", IEEE J. Solid-State Circuits, Vol SC-4, pp.291-293, 1969.

[CHE77] P. C. CHEN, "Threshold-alterable Si-gate MOS devices", IEEE Trans. Elec. Devices, Vol ED-24, p.584, 1977.

[CHE93] J.C. CHEN et al., "Degradation of N2O-annealed MOSFET characteristics in response to dynamic oxide stressing", IEEE Electron Device Lett., Vol. 14, no 5, p.225, 1993.

[CHO94] W. L. CHOI, D. M. KIM, "A new technique for measuring coupling coefficients and 3 D capacitance characterization of floating gate devices", IEEE Trans. Electron Devices, Vol 41, no 12, pp.2337-2342, 1994.

[CRR96] C. de GRRAF *et al.*, "A novel high-density low cost diode programmable read only memory", proc. IEEE International Electron Meeting (IEDM), p.89, 1996.

[DAG00] J.M. DAGA, C. PAPAIX, M. MERANDAT, S. RICARD, G. MEDULLA, J. GUICHAOUA, D. AUVERGNE, "Design techniques for embedded EEPROM memories in portable ASIC and ASSP solutions", Memory Technology, Design and Testing, Records of the 2000 IEEE International Workshop on, pp.39-447-8, 2000.

[DAG02] J.M. DAGA, "Test and repair of embedded flash memories", Test Conference, Proceedings. International, pp.1219, 2002.

[DAG03] J.M. DAGA, C. PAPAIX, M. MERANDAT, S. RICARD, G. MEDULLA, J. GUICHAOUA, D. AUVERGNE, "Design techniques for EEPROMs embedded in portable systems on chips", Design & Test of Computers, IEEE , Vol 20 , Issue: 1, pp.68-75, 2003.

[DAV89] R. DAVID, A. FUENTES, B. COURTOIS, "Random Pattern Testing versus deterministic testing of RAM's", IEEE Transactions on Computers, Vol C-38, no 5, pp. 637-650, 1989.

[DEJ76] J.H. DE JONGE, A.J. SMEULDERS, "Moving Inversions Test Pattern is Thorough, Yet Speedy", Computer Design, pp.169-173, 1976.

[DEK90] R. DEKKER *et al.,* "A Realistic Fault Model and Test Algorithm for Static Random Access Memories", IEEE Transactions on Computer, Vol C-9(6), pp.567-572, 1990.

[DIC76] J. DICKSON, "On chip high voltage generation in MNOS integrated circuits using an improved voltage multiplier technique", IEEE J. Sol. St. Cir., Vol.11, p.374, 1976.

[DOE78] D.H. Doehlert, "Uniform Shell Designs", Applied Statistics, Vol 19, pp.231-239, 1978.

[DRG96] R. DEGRAEVE *et al.*, "A new polarity dependence of the reduced trap generation during high-field degradation of nitrided oxides", IEDM Tech. Dig., p.327, 1996.

[ECS02] Rapport interne STmicroelectronics ECS17502, "Evaluation des résistances de source : technologie F6DP 5.2 μm^2", laboratoire électrique, Rousset, France, 2002.

[EIT96] B. EITAN *et al.*, "Multilevel flash cells and their trade-offs", proc. IDEM, p.169, 1996.

[ELD98] "ELDO User's Manual", Société Mentor Graphics Corp.,1998.

[FAL91] M. FALLON, M. ROBERTSON, A.J. WALTSON, R.J. HOLWILL, "Examination of LOCOS process parameters and the measurement of effective width", Microelectronic Test Structures, ICMTS'91, Proceedings of the 1991 International Conference on, pp.157-161, 1991.

[FAN98] Y.-H. FAN, Y. MOALEM, "Effective defect detection and classification methodology based on integrated laser scanning inspection and automatic defect classification", Advanced Semiconductor Manufacturing Conf. and Work., IEEE/SEMI, pp.266-271, 1998.

[FOT97] D. FOTY, "MOSFET Modelling with SPICE, Principles and Practice", Prentice Hall, ISBN 0-13-227935-5, 1997.

[GAL78] J. GALIAY, "Conception de circuits à large échelle d'intégration facilement testables", thèse de Doctorat, Université Paul Sabatier, Toulouse, 1978.

[GHI99] H.E. GHITANI, "DIBL coefficient in short-channel NMOS transistors", Radio Science Conference, NRSC '99, Proceedings of the Sixteenth National, pp. D4/1-D4/5, 1999.

[GRI96] G. CRISENZA, R. ANNUNZIATA, E. CAMERLENGHI, P. CAPPELLETTI, "Non volatile memories: Issues, challenges and trends for the 2000's scenario", Proc. ESSDERC'96, pp.121-130, 1996.

[HAK97] K. HAKOZAKI, S.-I SATO, K. IGUCHI, K. SAKIYAMA, "A new technique and a test structure for evaluating V_{th} distribution of flash memory cells", Proceedings of IEEE International Conference on Microelectronic Test Structures, pp127-130, 1997.

[HAR78] E. HARIRI, L. SCHMITZ, B. TROUTMAN, S. WANG, "A 256 bit non-volatile static RAM", IEEE ISSCC Dig. Tech. Pap., p.108, 1978.

[HAR92] R.E. HARRIS, "Defect density assessment in an integrated circuit fabrication line", Defect and Fault Tolerance in VLSI Systems, Proceedings., IEEE International Workshop on, pp.2-11, 1992.

[HAR00] N. HARABECH, R. BOUCHAKOUR, P. CANET, Ph. PANNIER, J.P. SORBIER, "Extraction of Fowler-Nordheim Parameters of thin SiO2 oxide film including polysilicon gate depletion: validation with an EEPROM memory cell", conf. IEEE ISCAS, Genève, 2000.

[HAY00] R. HAYTHORNTHWAITE, "Failure Mechanisms in Semiconductor Memory Circuits", Memory Technology, Design and Testing, 2000. IEEE Int. Workshop, pp.7-13, 2000.

[HIM95] T.HIMENO *et al.*, "A New Technique for Measuring Threshold Voltage Distribution in Flash EEPROM Devices", Proc. IEEE Int. Conf. of Microelectronic Test Structures, pp.283-287, 1995.

[HUF80] H. R. HUFF, R. D. HALVORSON, T. L. CHIU, D. GUTERMAN, "Experimental observations on conduction through polysilicon oxide", J. Electrochem. Soc., Vol 127, p.2482, 1980.

[JEE93] A. JEE *et al.*, "Carafe: A Software Tool for Failure Analysis", Proc. Of Int'l Symp. On Testing and Failure Analysis, pp.143-149, 1993.

[JEX91] J. JEX, A. BAKER, "Content addressable memory for flash redundancy", Communications, Computers and Signal Processing, IEEE Pacific Rim Conference on, Vol 2, pp.741-744, 1991.

[KAH67] D. KAHNG, S. M. SZE, "A floating gate and its application to memory devices", Bell Syst. Tech. J., Vol 46, pp.1288, 1967.

[KHU97] R. KHUBCHANDANI, "A new technique and a test structure for evaluating V_{th} distribution of flash memory cells", Proc. IEEE Microelectronic Test Structures, pp.1-6, 1997.

[KLE79] R. KLEIN, W. OWEN, R. SIMKO, W. TCHON, "5-V, non volatile RAM owes it all to polysilicon", Electronics, p.111, 1979.

[KOB95] SHIN-ICHI KOBAYASHI *et al*, "A 3.3V-Only 16 Mb DINOR Flash Memory", IEEE Solid-State Circuits Conference, 1995.

[KOR01] I. KOREN, "Yield: Statistical Modeling and Enhancement Techniques", Yield Optimization and Test Workshop (YOT'01), 2001, Internet source: http://www.ecs.umass.edu/ece/koren/yield/yotslides.pdf.

[KOV81] P. KOVIJANIC, "Single testability figure of merit", IEEE International test conference, Philadelphia, 1981.

[KUM01] S. KUMAR, E.L. RUSSEL, "Model to predict reliability of ONO non-volatile memory", Integrated Reliability Workshop Final Report, IEEE International, pp.34-40, 2001.

[LAN99] C. LANDRAULT, "Test, testabilité, et test intégré des circuits intégrés logiques", Cours, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRM), 1999.

[LAN80] G. LANDERS, "5-V only EEPROM mimics static RAM timing", Electronics, p.127, 1980.

[LEE97] A. LEE, L. MILOR, LIN YUNG-TAO, "The optimization of in-line scanner defect sizing using a circuit's layout and critical area", Advanced Semiconductor Manufacturing Conference and Workshop, 1997, IEEE /SEMI, pp.78-83, 1997.

[LEP94] D. LEPEJIAN *et al*, "An Automated Failure Analysis (AFA) Methodology for Repeated Structures", Proc. 12th IEEE VLSI Test Symp., IEEE Computer Society Press, Los Alamitos, Calif., pp.319-324, 1994.

[LIU98] W. LIU *et al*, "BISIM3v3 MOSFET model - User's manual", Univ. Of California, Berkeley, 1998, Internet source: http://www-device.eecs.berkeley.edu/~bsim/.

[MAN99] Y. MANEGLIA, D. BAUZA, "Evolution of the Si-SiO2 interface trap characteristics with Fowler-Nordheim injection", Microelectronic Test Structures, ICMTS, pp.117-120, 1999.

[MAS00] P. MASSON, J-L. AUTRAN, "transistor MOS à effet de champ, éléments de théorie et de pratique", Cours Université Aix-Marseille I, Laboratoire de Matériaux et de Microélectronique de Provence (L2MP), Edition 2000-2001.

[MAT99] D. MATHIEU, J. NONY, R. PHAN-TAN-LUU, "Logiciel NEMROD : Génération de matrices d'expériences en fonction des objectifs et traitement des réponses expérimentales", version 9901, LPRAI, Marseille, France, 1999.

[MAZ92] P. MAZUNDER, J.H. PATEL, "An efficient design of embedded memories and their testability analysis using Markov chains", Journal of Electronic Testing: Theory and Applications, Vol 3, pp. 235-250, 1992.

[MIE87] N. MIELKE, A. FAZIO, H. C. LIOU, "Reliability comparison of FLOTOX and textured polysilicon EEPROM's", IRPS Proc. Int. Rel. Phys. Symp., p.85, 1987.

[MIL99] L.S. MILOR, "Yield modeling based on in-line scanner defect sizing and a circuit's critical area", Semiconductor Manufacturing, IEEE Transactions on, Vol 12, pp.26-35, 1999.

[OHM94] K. OHMI *et al.*, "Hydrogen radical balanced steam oxidation for growing ultra-thin high reliability gate oxide films", Symp. VLSI Technol. Dig. Techn. Pap., p.109, 1994.

[OTT99] R. OTT *et al*, "An Effective Method to Estimate Defect Limited Yield Impact on Memory Devices", Proceedings IEEE/SEMI Advanced Semiconductor Manufacturing Conference, pp.87-91, 1999.

[PAL00] L. PALUN, "Etude prospective sur les dispositifs silicium à blocage de coulomb dans une perspective d'application à la micro électronique", Thèse, Université Joseph Fourrier, 2000.

[PAP85] C.A. PAPACHRISTOU, N.B. SAGAL, "An Improved Method for Detecting Functional Faults in Random–Access Memories", IEEE Transactions on Computers, Vol C-34, no 2, pp.110-116, 1985.

[PET61] W. WESLEY PETERSON, "Error-Correcting Codes", 1961.

[PHI00] Philips Semiconductors, "The I²C - bus specification", document order number: 9398 393 40011, Version 2.1 January 2000, Internet source: http://www.semiconductors.philips.com/buses/i2c/.

[PHI01] Philips Semiconductors, MM9 Model, 2001, Internet source : http://www.semiconductors.philips.com/philips_model/mos_models/.

[POR02] J.M. PORTAL, L. FORLI, H. AZIZA, D. NEE, "An Automated Methodology to Diagnose Geometric Defect in the EEPROM Cell", Proceedings of the IEEE International Test Conference, USA, pp.31-36, 2002.

[POR02] J.M.PORTAL *et al.*, "Floating-Gate EEPROM Cell Model Based on MOS Model 9", IEEE Int'l Symp. on Circuits and Systems, 2002.

[PRA87] K. PRALL, W. I. KINNEY, J. MARCO, "Characterization and suppression of drain coupling in submicrometer EPROM cells", IEEE Trans. Electron Devices, Vol ED-34, no 12, p.2463, 1987.

[REN86] M. RENOVELL, "Contribution à la modélisation des fautes dans les circuits intégrés MOS", Thèse de doctorat, LAMM–USTL, Montpellier, 1986.

[ROC94] M. ROCA, F. MOLL, A. RUBIO, "Crosstalk effects between metal and polysilicon lines in CMOS integrated circuits", IEEE transactions on EMC, Vol 36, no 3, pp.250-253, 1994.

[SAK90] K. SAKALLAH *et al*, "A First-Order Charge Conservative MOS capacitance Model", IEEE Trans. Comp.-Aid. Des., Vol 9, pp.99-108, 1999.

[SAK94] K. SAKUI, F. MASUOKA, "Sub-halfmicron flash memory technologies", Proceedings of IEICE Trans. Electron., Vol E77-C, no 8, pp.1251-1259, 1994.

[SAL99] B. DE. SALVO, "Etude du transport électrique et de la fiabilité des isolants mémoires non volatiles à grille flottante", Thèse, Institut National Polytechnique de Grenoble, 1999.

[SAM95] P. SAMANTA, C. K. SARKAR, "Influence of neutral hole traps in thin gate oxides on MOS device degradation during Fowler-Nordheim stress", IEEE Region Microelectronics and VLSI, pp.260-262, 1995.

[SEG99] J. SEGAL *et al*, "Determining redundancy requirements for memory arrays with critical area analysis", Memory Technology, Design and Testing, Records of the 1999 IEEE International Workshop on, pp.48-53, 1999.

[SEG00] J. SEGAL *et al*, "Critical area based yield prediction using in-line defect classification information", Advanced Semiconductor Manufacturing Conference and Workshop, IEEE/SEMI, 2000.

[SEG01] J. SEGAL *et al*, "Using Electrical bitmap results from Embedded memory to enhance yield", IEEE Design & Test of Computer, Vol 18, no 3, pp.28-39, 2001.

[SIC98] E. SICARD, "Le Couplage Diaphonique dans les Circuits CMOS Sub-microniques", Thèse, Département de génie électrique et informatique, Institut National des Sciences Appliquées, Toulouse, France, 1998.

[SMF85] J.P. SHEN, W.MALY, F.J. FERGUSON, "Inductive fault analysis of MOS integrated circuits", IEEE Design and Test of Computers, pp.13-26, 1995.

[SOB95] SIK-HAN SOB, C. MESSICK, CHIH-CHANG CHEN, CHUAN-PU LIU, B. BAKEL, R. PRTERSON, L. SADWICK, "Fast test at wafer-level for endurance of tunnel oxide", Integrated Reliability Workshop, Final Report., International, 1995.

[SUZ83] E. SUZUKI, H. HIRAICHI, K. ISHI, Y. HAYASHI, "A low voltage alterable EEPROM with metal-oxide-nitride-oxide-semiconductor (MNOS) structure", IEEE Trans. Elec. Devices, Vol ED-30, p.122, 1983.

[TAK97] M. TAKEO, M. AZUMA, T. SUMI, K. TATSUMA, "Integrated test circuit to measure polarisation characteristics of ferroelectric capacitors for development of Mega-Bit scale Feram", IEE Proc. Microelectronics Test Structures ICMTS, 1997.

[TAM67] E. TAMMARU, J. B. ANGELL, "Redundancy for LSI yield enhancement", IEEE J. Solid-State Circuits, Vol SC-2, pp.172-182, 1967.

[TAN99] T. TANZAWA *et al.*, "Optimization of Wordline Booster Circuits for Single Power Supply Flash Memories", IEEE J. of Solid States Circuits, Vol 32, no 8, pp.1091-1098, 1999.

[TOY80] H. TOYOSHIMA *et al*, "Feram device and circuit technologies fully compatible with advanced CMOS", IEEE conf. Custom integrated Circuits, 2001.

[UYE97] J.P. UYEMURA, "Circuit Design for CMOS VLSI", Kluwer Academic Publishers, 1997.

[VAN90] A.J. VAN DE GOOR, C.A. VERRUIJT, "An overview of Deterministic Functional RAM Chip Testing", ACM Computing Surveys, 22(1), pp.5-33, 1990.

[VAN91] A.J. VAN DE GOOR, "Testing Semiconductor Memories", John Wiley & Sons, 1991.

[VAN93] A.J. VAN DE GOOR, "Using March Tests to Test SRAM", IEEE Design & Test of Computers, pp.8-14, March 1993.

[WAD78] R.L. WADSACK, "Fault Modeling and Logic Simulation of CMOS and MOS Integrated Circuits", The Bell System Technical Journal, Vol 57, no. 5, 1978.

[WAD80] M. WADA, S. MIMURA, H. NIHIRA, H. IZUKA, "Limiting factors for programming EPROM of reduced dimensions", IEDM Tech. Dig., pp.38-41, 1980.

[WAI90] J.A. WAICUKAUSKI, P.A. SHUPE, D.J. GIRAMMA, A. MATIN, "ATPG for ultra-large structured designs", Test Conference, Proceedings, International, pp.44-51, 1990.

[WAL94] J. WALKER, D. T. CHI, "Forward Error Correction for Solide-State Memory Array", NML Spring Conference, Appendix, pp.111-120, 1994.

[WEG67] H. R. WEGENER, A. J. LINCOLN, H. C. PAO, M. R. O'CONNEL, R. E. OLEKSIAK, "The variable threshold transistor, a new electrically alterable, non-destructive read-only storage device", IEEE IEDM Tech. Dig., 1967.

[WEL95] D. WELLEKENS *et al.*, "Write/Erase Degradation in Source Side Injection Flash EEPROM's: Characterization Techniques and Wearout Mechanisms", IEEE Transactions On Electron Devices, Vol 42, no 11, pp.1992-1998, 1995.

[WIC99] G. WICKER, "Nonvolatile, high density, high performance phase change memory", SPIE, Vol 3891, pp.2-9, 1999.

[WIL94] D. WILSON, A.J. WALTON, "Automatic in-line measurement for the identification of killer defects", Semiconductor Processing Quality through Measurement, IEE Colloquium on, pp.5/1-5/8, 1994.

[WON92] M. WONG, D. K.-Y. LIU, S. S.-W. HUANG, "Analysis of the subthreshold slope and the linear transconductance techniques for the extraction of the capacitive coupling coefficients of floating-gate devices," IEEE Electron Device Lett., Vol 13, no 11, pp.566-568, 1992.

[WU83] T. C. WU, W. T. STACY, K. N. RITZ, "The influence of the LOCOS Processing Parameters on the shape of the Bird's Beak Structure", J. Electrochem. Soc., Vol 130, no 7, pp.1563-1566, 1983.

[YAR92] G. YARON, S. PRASARD, M. EBEL, B. LEONG, "A 16K EEPROM employing new array architecture and new designed-in reliability features", IEEE J. Sol. St. Circuit, Vol SC-17, no.5, p.933, 1992.

Valorisation scientifique

Publications	Nombres
Revues	1
Conférences Internationales avec actes	4
Communications sans actes	5
Brevets	2

Revues :

Portal J.M., Aziza H., Née D.
« EEPROM Diagnosis Based on Threshold Voltage Embedded Measurement »
Journal of Electronic Testing (JETTA), accepté pour publication

Conférences Internationales avec actes et comités de lecture :

Portal J.M., Aziza H., Née D.
« EEPROM Memory Diagnosis Based on Threshold Current Extraction »
Proceedings of the IEEE Conference on Design of Circuits and Integrated Systems, Espagne, 2003, pp.133-139

Portal J.M., Aziza H., Née D.
« EEPROM Memory: Threshold Voltage Built In Self Diagnosis »
Proceedings of the IEEE International Test Conference, USA, 2003, pp.23-28

Portal J.M., Forli L., Aziza H., Née D.
« An Automated Methodology to Diagnose Geometric Defect in the EEPROM Cell »
Proceedings of the IEEE International Test Conference, USA, 2002, pp.31-36

Portal J.M., Forli L., Aziza H., Née D.
« An Automated Design Methodology for EEPROM Cell (ADE) »
Proceedings of the IEEE International Workshop on Memory Technology Design and Testing, France, 2002, pp.137-142

Communications :

Aziza H., Portal J.M., Née D.
« EEPROM Threshold Current Extraction: Silicon Validation »
Proceedings of the IEEE European Test Symposium, France, 2004, pp.81-87

Portal J.M., Aziza H., Née D.
« EEPROM Memory: Threshold Voltage Built In Self Diagnosis »
Proceedings of the IEEE European Test Workshop, Hollande, 2003, pp.81-87

Portal J.M., Aziza H., Née D.
« Mémoires EEPROM : Extraction des tensions de seuil en vue du diagnostic »
Journées Nationales du Réseau Doctoral de Microélectronique, France, 2003

Portal J.M., Forli L., Aziza H., Née D.
« An Automated Geometric Defect Diagnosis Methodology for EEPROM Cell (AGDE) »
Proceedings of the IEEE European Test Workshop, Grèce, 2002, pp.343-344

Aziza H., Portal J.M., Née D.
« Modèle de cellule EEPROM base sur le MOS modèle 9 »
Journées Nationales du Réseau Doctoral de Microélectronique, France, 2002

Brevets :

Portal J.M., Aziza H., Née D.
« Circuit de caractérisation des tensions de seuils dans les mémoires non volatiles sous format numérique et traitement en vue du diagnostic des données obtenues »
N° de dépôt 0306184, N° de dossier 03-RO-059, 2003

Portal J.M., Forli L., Aziza H., Née D.
« Procédé de modélisation mathématique de composants électroniques et son utilisation pour la simulation, la détermination de géométrie et le diagnostic de défauts dans le composant »
N° de dépôt 0208652, N° de dossier 02-RO-029, 2002

Glossaire

ASIC (Applied Specific Integrated Circuits) : Circuits intégrés pour applications spécifiques.

ATPG (Automatic Test Pattern Generator) : Outil de génération automatique de vecteurs de test.

ATE (Automated Test Equipement) : Equipement destiné au test de circuits VLSI, communément appelé testeur.

Back-end : Assemblage et essai. Dans l'industrie des semiconducteurs, le Back-End correspond à la seconde phase de la fabrication au cours de laquelle le circuit en silicium est monté dans un boîtier (assemblage) conçu non seulement pour le protéger, mais aussi pour assurer des connexions avec l'extérieur par le biais de fils très fins. Cette opération est suivie des opérations suivantes : essais, assemblage, finition et emballage. Dans le cadre de cette thèse le procédé de fabrication Back-End regroupe toutes les étapes de fabrication liées aux interconnexions.

Binning : Opération de tri qui permet de séparer les produits testés en différentes catégories définies en fonction de leurs performances.

BIST (Built in self Test) : Techniques de test destinées à inclure directement dans le circuit tout ou partie des fonctions réalisées par le testeur.

Bit : Valeur numérique telle que 0 ou 1. Ou toute valeur unique à deux états (On/Off, Vrai/Faux).

Bit line : Colonne d'une matrice de cellules mémoire.

Bonding : Processus de soudage dans lequel les connexions sont établies avec les contacts de surface d'un circuit intégré par soudage par thermocompression avec des fils d'or.

Burn-in (Déverminage) : Essai à haute température utilisé pour détecter et éliminer tout défaut pouvant apparaître pendant la prime jeunesse d'un produit.

CAO (Conception Assistée par Ordinateur) : Ensemble d'outils matériels et logiciels utilisés pour la conception graphique. Facilitent la conception d'un produit et la vérification de ses performances par simulation.

Carte à pointe (Probe Card) : Carte équipée de minuscules aiguilles ou sondes qui établissent le contact avec les « pads » (entrées-sorties) des puces sur la plaquette de silicium.

Circuit intégré (Integrated circuit - IC) : Contrairement aux dispositifs discrets, les circuits intégrés incorporent plusieurs fonctions sur la même puce. Suivant le degré d'intégration, on parlera de circuits intégrés SSI (Intégration à faible échelle), MSI (intégration à moyenne échelle), LSI (intégration à grande échelle), VLSI (intégration à très grande échelle) ou ULSI (intégration à ultra-grande échelle). Un circuit intégré peut héberger sur la même puce quelques transistors (plus les diodes, les résistances et les condensateurs nécessaires pour réaliser un circuit complet), mais aussi plusieurs millions de transistors.

CMOS (Complementary Metal Oxide Semiconductor) : Technologie utilisée pour concevoir et fabriquer des transistors NMOS (transistors MOS dont le flux de charges est négatif) et des transistors PMOS (transistors MOS dont le flux de charges est positif) sur un même substrat. Cette technologie est réputée pour sa faible consommation.

CPU (Central Processing Unit) : Unité centrale.

Data-sheet : Spécifications du constructeur pour un produit donné.

Dépôt en phase vapeur (Chemical vapor deposition - CVD) : Processus utilisé pour déposer une pellicule mince à la surface d'une tranche de silicium. Il existe deux façons d'effectuer un dépôt : dépôt au plasma ou à basse pression.

Dépôt épitaxial (Epitaxial deposition) : Dépôt d'une seule couche de cristal sur un substrat de sorte que la structure cristalline de la couche déposée corresponde à la structure cristalline du substrat.

DIB (Device Interface Board) : Carte qui constitue l'interface physique entre le système de test et le produit testé. Elle contient les composants de l'interface (relais, résistances ou capacités) nécessaires au test de la puce.

Diffusion (Diffusion) : Injection d'éléments dopants spécifiques dans la structure de crystal de la tranche de silicium.

DOE (Design Of Experiment) : Méthodologie d'analyse qui permet de déterminer, grâce à une série d'expérimentations judicieusement choisies, les facteurs ou paramètres ayant une influence significative sur une caractéristique à maîtriser (réponse). Ainsi, il est possible d'arriver à la connaissance de la réponse pour n'importe quel point du domaine expérimental défini par les facteurs d'entrée.

DPS (Device Power Supply) : Les DPS fournissent la tension et le courant aux broches d'alimentation de la puce.

DRAM (Dynamic RAM) : Les mémoires DRAM permettent d'accéder individuellement à chaque cellule pour écrire, lire ou ré-écrire des données autant de fois que nécessaire. Les données sont conservées tant que la mémoire est alimentée électriquement. Un circuit de commande rafraîchit la charge dans chaque cellule à une fréquence très élevée. Les mémoires DRAM sont utilisées dans tous les secteurs du traitement électronique des données et les micro-ordinateurs en sont la principale application.

DUT (Device Under Test) : Circuit à tester.

EEPROM (E^2PROM) : Les mémoires EEPROM (Electrically Erasable Programmable ROM) sont effaçables électriquement par unité d'information élémentaire. Les EEPROM remplacent les EPROM lorsqu'une reprogrammation interne est nécessaire.

EPROM : Les mémoires EPROM (Erasable programmable Read Only Memory) sont programmées à l'aide de signaux électriques et peuvent être effacées électriquement par rayons ultraviolets avant d'être reprogrammées. Les mémoires EPROM conviennent idéalement au stockage des données fixes ou lorsque des programmes standard sont nécessaires.

EWS (Electrical Wafer Sort) : Tri électrique des tranches de silicium.

Fab : Usine de fabrication. Une usine de fabrication nécessite un environnement particulier. Les critères de propreté requis pour les processus haute précision sont extrêmement sévères. L'air des

salles de fabrication est de 10.000 à 100.000 fois plus pur que l'air ambiant et les opérateurs portent des tenues spéciales.

Flash : Les mémoires Flash associent la haute densité et l'excellent rapport coût/performances des mémoires EPROM à la possibilité d'effacement électrique des mémoires EEPROM. C'est pourquoi le marché des mémoires flash est aujourd'hui l'un des plus intéressants pour l'industrie des semiconducteurs. On distingue deux catégories de mémoires Flash : les mémoires NOR et NAND. Elles utilisent des architectures distinctes, dont les avantages en termes d'applications diffèrent totalement. La cellule mémoire NAND est environ 40% plus petite qu'une cellule NOR, ce qui se traduit par un coût au bit moins élevé.

Front-end : Dans l'industrie des semiconducteurs, le Front-end correspond à la première phase de la fabrication des puces sur silicium, juste avant leur mise en boîtier. Dans le cadre de cette thèse le procédé de fabrication Front-End regroupe toutes les étapes liées à la fabrication des transistors.

Handlers : machines qui permettent de manipuler les puces mises en boîtier.

HV(OX) : Oxyde haute tension.

LSI (Large Scale Integration) : Intégration à grande échelle.

LOCOS (LOCal Oxidation of Silicon) : Oxyde épais encore appelé oxyde de champ d'une épaisseur de l'ordre de 600 nm. Cet oxyde enterré permet d'isoler tous les composants afin d'éviter les fuites de courant entre les transistors adjacents.

MCU *:* Microcontrôleur.

Métallisation (Metalization) : Dépôt d'une mince couche de métal sur une tranche afin de permettre l'interconnexion des éléments d'un circuit intégré.

MOS (Metal Oxide Semiconductor) : L'une des technologies de base utilisées pour fabriquer des circuits intégrés et dans laquelle les charges électriques sont portées par conduction.

MOSFET (Metal Oxide Semiconductor Field Effect Transistor) : Transistor à effet de champ réalisé en technologie MOS.

MSI : Intégration à Moyenne Echelle.

NMOS : Transistor MOS à canal N où le flux de charge est négatif.

Non Volatilité : Dans une mémoire non volatile, les données mémorisées sont conservées même en cas de mise hors tension.

NVRAM (Non Volatile RAM) : Mémoire vive non volatile.

OTP (One Time Programmable) : Mémoire non volatile programmable une seule fois, comme une EPROM.

Pattern Memory : Chaque système de test possède une mémoire à accès rapide appelée « Pattern Memory » ou « Vector Memory », pour stocker les vecteurs de test (ou patterns de test). Les patterns de test contiennent les états des entrées et des sorties pour différentes fonctions logiques que le produit doit remplir. Au cours du test, les patterns sont appliqués sur la puce par le système de test et les signaux de réponse sont capturés. Si la réponse attendue ne correspond pas à la réponse issue du produit testé, un échec fonctionnel survient.

Photolithogravure (Photolithography) : Processus au sein duquel le motif représentant les composants d'un circuit intégré est transposé sur une tranche au moyen de la lumière.

Photomasque (Photomask) : Support utilisé pour transférer une image sur une tranche pendant la fabrication de celle-ci. Un photomasque est réalisé en verre et revêtu de chrome.

Photorésist (Photoresist) : Matériau sensible à la lumière utilisé pendant le processus de photolithogravure. Aussi appelé résine.

PLA (Programmable Logic Array) : Réseau logique programmable.

PMOS : Transistor MOS à canal P où le flux de charge est positif.

PMU (Precision Measurement Unit) : Le PMU est utilisé pour effectuer des mesures de tension et courant précises. Il peut forcer un courant et mesurer une tension ou forcer une tension et mesurer un courant. Une sélection d'échelle ou « range » appropriée doit assurer les résultats de test les plus précis possibles. Le PMU possède deux limites de mesure programmables, une limite supérieure et une limite inférieure. Ces limites peuvent être utilisées individuellement ou simultanément. Si la valeur mesurée n'appartient pas à ces intervalles, le test a échoué (FAIL). Si la valeur mesurée est comprise dans les limites de test définies, le test est réussi (PASS).

ppm (parts per million) : Abréviation utilisée pour indiquer le nombre de parties étrangères ou différentes d'un ensemble. Dans le domaine de la qualité, cette abréviation indique le nombre de défauts.

Prober : Machines qui manipulent les plaquettes de silicium durant le test de production.

PROM (Programmable Read Only Memory) : Type de mémoire non volatile programmable électriquement.

Puce (chip) : Terme désignant un circuit à semiconducteur.

R&D (Research & Development) : Recherche & Développement.

RAM (Random Access Memory) : Mémoire vive. Les premières mémoires d'ordinateurs disposaient d'un accès série. Les mémoires dont toutes les adresses sont accessibles à tout moment étaient appelées « mémoires vives » par opposition aux mémoires dont le contenu est accessible uniquement selon un ordre donné. Ce terme est utilisé aujourd'hui pour désigner les mémoires vives non volatiles à semiconducteurs.

Rendement (Yield) : Lors de l'opération de tri électrique des tranches, le rendement désigne les puces bonnes (portion électrique de la tranche contenant les fonctions électroniques) par rapport au nombre total de puces sur la tranche.

ROM (Read Only Memory) : Mémoire morte (non volatile).

RVS (Reference Voltage Supply) : Les RVS fournissent les tensions de référence pour les niveaux logiques 0 et 1. Ces tensions sont représentées par les symboles V_{IL}, V_{IH}, V_{OL} et V_{OH}.

SEM (Scanning Electron Microscope) : Equipement utilisé pour analyser les défauts.

SRAM (Static RAM) : Les mémoires SRAM présentent les mêmes fonctions que les mémoires RAM dynamiques (DRAM) mais en offrant une plus grande rapidité sans nécessiter de rafraîchissement permanent. Malheureusement, les SRAM nécessitent une alimentation électrique plus puissante. Les cellules des SRAM sont plus complexes que celles des DRAM, ce qui explique que les SRAM offrent

un quart de la capacité de stockage de leurs équivalents DRAM les plus récents pour un coût supérieur. Les mémoires SRAM sont souvent utilisées en petites quantités comme mémoires cache dans les microordinateurs. Elles permettent de stocker des données fréquemment utilisées par le microprocesseur, ce qui assure un accès aux données plus rapide que la mémoire principale DRAM.

TEG (Test Element Group) : Motif de test. Généralement inséré dans les lignes de découpe des tranches de silicium.

ULSI (Ultra-Large Scale Integration) : Intégration à ultra-grande échelle.

VLSI (Very Large Scale Integration) : Intégration à très grande échelle.

Word line : Ligne de mot, correspond à la ligne de sélection d'un mot.