THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : PHYS - Physique
Spécialité : Nanophysique
Unité de recherche : CEA LETI - Laboratoire d'électronique et de technologie de l'information

Conception de circuits intégrés pour l'amélioration des performances des mémoires OxRAM

Integrated circuits to enhance the performance of resistive-based memories

Présentée par :

Thomas BAUVENT

Direction de thèse :

Gaël PILLONNET DIRECTEUR DE RECHERCHE, CEA CENTRE DE GRENOBLE	Directeur de thèse
Gabriel MOLAS SENIOR SCIENTIST, WEEBIT NANO FRANCE	Co-directeur de thèse

Rapporteurs :

MARC BOCQUET
PROFESSEUR DES UNIVERSITES, AIX-MARSEILLE UNIVERSITE
LIONEL TORRES
PROFESSEUR DES UNIVERSITES, UNIVERSITE DE MONTPELLIER

Thèse soutenue publiquement le **15 décembre 2025**, devant le jury composé de :

LORENA ANGHEL, PROFESSEURE DES UNIVERSITES, GRENOBLE INP - UGA	Présidente
MARC BOCQUET, PROFESSEUR DES UNIVERSITES, AIX-MARSEILLE UNIVERSITE	Rapporteur
LIONEL TORRES, PROFESSEUR DES UNIVERSITES, UNIVERSITE DE MONTPELLIER	Rapporteur
ALEXANDRE LEVISSE, SENIOR SCIENTIST, ECOLE POLYTECHNIQUE FEDERALE DE LAUSANNE	Examinateur

Résumé

Cette thèse traite de la conception de circuits mémoire utilisant l'OxRAM. Elle est focalisée sur deux problématiques, l'écriture des OxRAM et l'augmentation de la densité mémoire réalisable. Pour l'écriture, différents types de sources sont explorés. Les méthodes d'écriture utilisant des sources de tension, très présentes dans la littérature, servent de base de comparaison. Des méthodes d'écriture utilisant des sources de courant sont étudiées, d'abord à l'aide d'un premier circuit intégré issu d'une précédente thèse, puis d'un circuit intégré spécifiquement développé dans le cadre de cette thèse. Ce dernier permet une comparaison au sein d'un même circuit entre les méthodes d'écriture utilisant des sources de tension et celles utilisant des sources de courant. Différents algorithmes de programmation sont considérés dans l'objectif de réduire la variabilité de l'écriture. Des méthodes d'écriture utilisant des sources capacitives sont également discutées, et un circuit permettant leur étude est présenté, cependant sans résultats expérimentaux. Enfin, pour la densité, une nouvelle architecture de matrice 1TnR n'utilisant qu'un sélecteur pour n éléments OxRAM est développée d'abord sur un plan théorique, puis dans une implémentation intégrée.

Remerciements

Cette thèse n'aurait probablement pas eu lieu sans l'aide de nombreuses personnes que je souhaiterais remercier ici.

En commençant par ceux ayant marqué mon parcours m'ayant permis de débuter cette thèse :
- Merci à Claude Chappert et Thibaut Devolder pour m'avoir permis d'avoir mon premier contact avec le milieu des nanosciences au travers d'un stage de première année de master dans le Novel Magnetic Devices Group du Centre de Nanosciences et de Nanotechnologies en manipulant des STT-MRAM.
- Merci à Bruno Le Pioufle et Yoshio Mita (三田 吉郎) pour m'avoir donné l'opportunité de découvrir le Japon et la microélectronique au travers d'un stage de 9 mois dans le laboratoire de M. Mita (三田研究室) à l'Université de Tokyo (東京大学). Par la force des choses, ce fut également l'occasion de vivre le début de la crise internationale du COVID-19 d'une manière différente, contrastant avec mon année suivante en France.
- Enfin, merci à Gaël Pillonnet et Gabriel Molas pour m'avoir pris en stage de seconde année de Master au CEA-Leti, stage qui fut le prélude de cette thèse.

Merci à tous ceux qui ont participé de près ou de loin à la réussite de cette thèse lors de ces trois années et demi passées au CEA-Leti. Au sein de mon laboratoire de rattachement, merci à François Andrieu et Gabriel Molas pour avoir financé cette thèse. Au sein de mon laboratoire d'accueil, merci à Stéphanie Robinet pour l'accueil dans son laboratoire. Merci à Yasser Moursy pour son encadrement et son engagement sur la partie électronique analogique de cette thèse. Merci à Gaël Pillonnet pour la direction de thèse. Et merci à tous les autres membres du laboratoire pour leur accueil, leur aide et les bons moments passés ensemble.

Merci à Jonathan Blow pour avoir conçu et développé le langage de programmation « JAI » (faute de nom officiel) utilisé pour le développement de la plateforme de test dont nous discutons au chapitre Chapitre IV. Bien qu'en bêta, ce langage a grandement aidé et facilité le développement, en particulier grâce à son design précis et son confort d'utilisation hors pair.

Enfin merci à ma famille, sans qui rien de tout cela n'aurait été possible, et sans qui je ne serais pas là.

Table des matières

I Introduction 13
I.1 La mémoire 13
I.2 Mémoire résistive 16
I.3 Cas d'application 17
I.3.1 Mémoire embarquée 18
I.3.2 Accélération matérielle 18
I.3.2.a Memristors et logique 19
I.3.2.b Opération logique à la lecture 20
I.3.2.c Multiplication-accumulation analogique 21
I.4 Conclusion 22
Bibliographie 24
II Écriture de l'Oxram 27
II.1 Introduction 27
II.2 Éléments constituants et considérations générales 27
II.2.1 Structure du dispositif et description physique élémentaire 27
II.2.2 Cellule 1T1R, caractéristique quasi-statique et variabilités 29
II.2.3 Caractéristiques importantes d'une cellule 1T1R 32
II.3 Méthode d'écriture 34
II.3.1 Quasi-statique 35
II.3.2 Impulsion à tension constante à durée déterminée 35
II.3.3 Impulsion à tension constante à durée adaptée 36
II.3.4 Écriture avec vérification 36
II.3.4.a Ré-écriture avec vérification 37
II.3.4.b Impulsions adaptatives avec vérification 37
II.3.4.c Limitations 38
II.3.5 Forming 38
II.3.6 Type de source utilisée pour l'écriture 39
II.3.6.a Écriture avec une source de tension 39
II.3.6.b Écriture avec une source de courant 39
II.3.6.c Écriture avec une source capacitive 40
Bibliographie 41
III Écriture en courant sur des dispositifs unitaires 45
III.1 Introduction 45
III.2 Écriture en courant 46
III.2.1 La relation entre R_{LRS} et I_c^{eff} 46
III.2.2 Apparente équivalence entre sources avec limitation 46
III.2.3 Différences dues aux effets dynamiques 47
III.3 Circuit intégré *WT-PAOLA* 48
III.3.1 Intentions initiales du circuit 48
III.3.2 Structure 49
III.3.3 Utilisation effective du circuit 50
III.4 Résultats expérimentaux 50
III.4.1 Forming 50
III.4.2 Effet de l'amplitude du courant 51
III.4.3 Effet de la durée de l'impulsion 53
III.4.4 Vérification de l'endurance 55
III.4.5 Écriture avec vérification 56

III.5 Conclusion 59
Bibliographie 60
IV Écriture en courant sur une matrice mémoire 63
IV.1 Introduction 63
IV.2 Conception du circuit *MEMORIIC* 63
IV.2.1 Objectifs 63
IV.2.2 Architecture générale 65
IV.2.3 Source de courant 67
IV.3 Conception de la plateforme de test 68
IV.3.1 Partie matérielle 68
IV.3.2 Partie logicielle 70
IV.3.3 Opération sur les cellules et algorithmes d'écriture 73
IV.4 Résultats expérimentaux 74
IV.4.1 Formation du filament 74
IV.4.2 Écriture 81
IV.5 Conclusion 87
Bibliographie 88
V Écriture capacitive 89
V.1 Introduction 89
V.2 Prédiction théorique, motivations et risques 90
V.2.1 Principe de l'écriture capacitive 90
V.2.2 Contrainte d'intégration 91
V.2.3 Méthode pour la réduction de la capacité nécessaire 92
V.3 Mise en œuvre de l'écriture capacitive : Circuit *CAPA-WR* 93
V.3.1 Objectifs 93
V.3.2 Présentation du circuit *CAPA-WR* 95
V.3.3 Simulation 96
V.4 Conclusion 98
Bibliographie 98
VI Limites de l'augmentation de la densité 99
VI.1 Introduction 99
VI.2 Dispositif unitaire 100
VI.2.1 Intégration 100
VI.2.2 Lien entre surface, temps de formation et tension de formation 100
VI.3 Multiplexage des accès 101
VI.3.1 Adressage d'un dipôle 101
VI.3.2 Cellules 1S1R 104
VI.3.3 Cellules 1T1R 105
VI.4 Conclusion 107
Bibliographie 108
VII Une solution au problème de densité : les matrices 1TnR ? 109
VII.1 Introduction 109
VII.2 Matrice 1TnR 109
VII.2.1 Cellule 1TnR 109
VII.2.2 Intégration sous forme de matrice 110
VII.2.3 Technique d'écriture 112
VII.2.4 Technique de lecture 117
VII.2.4.a Lecture absolue du niveau de conductance 117
VII.2.4.b Lecture relative 117

VII.2.5 Choix du nombre n d'éléments mémoires OxRAM par sélecteur 120
VII.3 Mise en œuvre de l'architecture 1TnR : Circuit *OX1TnR* 121
VII.3.1 Objectifs 121
VII.3.2 Contraintes technologiques 122
VII.3.3 Conception des masques d'une matrice 1T8R 123
VII.3.4 Dérivation des autres matrices 128
VII.3.5 Motif pour la mesure des impédances des lignes 130
VII.4 Conclusion 132
Bibliographie 133
Conclusion 135
Bibliographie 143
Liste de publications 149
A Instructions disponibles sur la plateforme de tests du circuit *MEMORIIC* 150
B Détails du circuit *CAPA-WR* 152
B.1 Pads 152
B.2 Mémoire de configuration 153
B.3 Matrice mémoire 155
B.4 Mode écriture en tension 156
B.5 Mode de lecture 156
B.6 Méthode de mesure des capacités intégrées 156
B.7 Mode d'écriture capacitive sans recharge 157
B.8 Mode d'écriture capacitive avec recharge 157
C Preuve de la condition de séparabilité de $2^n - 1$ états avec la méthode de lecture relative sur une cellule 1TnR 158

Liste des acronymes

ADC	*Analog to Digital Converter*	Convertisseur analogique-numérique
ASIC	*Application Specific Integrated Circuit*	Circuit intégré spécialisé
CAD	*Computer Aided Design*	Conception assistée par ordinateur
CBRAM	*Conductive Bridge RAM*	RAM à pont conducteur
CEA	Commissariat à l'énergie atomique - Laboratoire d'électronique et de technologie de l'information	
CMOS	*Complementary MOSFET*	MOSFET complémentaire
DAC	*Digital to Analog Converter*	Convertisseur numérique-analogique
DRAM	*Dynamic RAM*	RAM dynamique
FDSOI	*Fully Depleted Silicon On Insulator*	Silicium complètement appauvri sur isolant
FPGA	*Field Programmable Gate Array*	Réseau de portes programmables
GPIO	*General Purpose Input/Output*	Entrée-sortie à usage général
HDD	*Hard Disk Drive*	Disque dur
HRS	*High Resistive State*	Etat de haute résistance
IDAC	*Current DAC*	DAC en courant
LRS	*Low Resistive State*	Etat de faible résistance
MAC	*Multiply-ACcumulate*	Multiplication-accumulation
MOM	*Metal Oxide Metal capacitor*	Capacité métal-oxyde-métal
MOS	*Metal Oxide Semiconductor capacitor*	Capacité métal-oxyde-semi-conducteur
MOSFET	*Metal Oxide Semiconductor Field Effect Transistor*	Transistor à effet de champ à grille métal-oxyde
MRAM	*Magnetic RAM*	RAM magnétique
NAND	*Not AND*	Opérateur logique non-et
NMOS	*N type MOSFET*	MOSFET de type N
NOR	*Not OR*	Opérateur logique non-ou
OTS	*Ovonic Threshold Switch*	Commutateur à seuil ovonique
OxRAM	*Oxide RAM*	RAM oxyde
PCM	*Phase Change Memory*	Mémoire à changement de phase
PMOS	*P type MOSFET*	MOSFET de type P
PWM	*Pulse Width Modulation*	Modulation de largeur d'impulsion
QLC	*Quad-Level Cell*	Cellule 4-bits
RAM	*Random Access Memory*	Mémoire à accès aléatoire
RC	Résistance-Capacité	
ReRAM	*Resistive RAM*	RAM résistive
RTN	*Random Telegraph Noise*	Bruit télégraphique
SMA	*SubMiniature version A*	Un type de connecteur coaxial
SMU	*Source Measure Unit*	Instrument de mesure pouvant mesurer le courant ou la tension tout en générant l'autre
SRAM	*Static RAM*	RAM statique
SSD	*Solid State Drive*	
STT-MRAM	*Spin Transfert Torque MRAM*	MRAM à transfert de spin
USB	*Universal Serial Bus*	Bus série universel

Liste des figures

Figure 2.1 Illustration des différents états et transitions entre ceux-ci dans un élément mémoire OxRAM. 27
Figure 2.2 Caractéristique quasi-statique d'une cellule 1T1R. Les flèches sur le cycle indiquent sa direction. Le temps de montée et descente de l'impulsion de tension est 350 kV/s (3,5 V en 10 µs) pour le SET et 180 kV/s (1,8 V en 10 µs) pour le RESET, avec un palier de 10 µs à la valeur maximale. 29
Figure 2.3 Illustration des relations entre les paramètres identifiés. T est l'ordre de grandeur du temps d'écriture et τ est le temps caractéristique de la réaction d'écriture. Les grandeurs sur fond bleu sont contrôlables par l'expérimentateur. Celles sur fond vert sont celles que l'on cherche à contrôler. Celles sur fond gris ne sont ni contrôlables, ni aisément observables. La flèche partant de (τ) indique que τ influence la relation entre I_c et I_c^{eff}. . 30
Figure 2.4 Caractéristique quasi-statique de quinze cellules 1T1R. Les conditions d'écriture sont identiques pour l'ensemble des quinze cellules et identiques aux conditions décrites dans la Figure 2.2. 30
Figure 3.1 Comparaison entre écriture en tension et écriture en courant en prenant en compte la capacité d'interconnexion C_p. 47
Figure 3.2 Schéma simplifié du circuit *WT-PAOLA*. 49
Figure 3.3 Dessin des masques (« layout » en anglais) du circuit *WT-PAOLA*. 49
Figure 3.4 Description de la procédure de formation des filaments utilisée (gauche). Diagramme de la procédure (droite). Exemple d'exécution de cette procédure. 50
Figure 3.5 Distribution cumulative normalisée du courant de lecture après un SET en courant pour différentes amplitudes d'impulsion. 51
Figure 3.6 Distribution cumulative normalisée du courant de lecture après un SET (rouge) et après un RESET (bleu) pour différentes longueurs d'impulsion de SET. 53
Figure 3.7 Nuage de points de la conductance de l'élément mémoire OxRAM au travers de 200 000 cycles, mesurée après le (rouge) SET, (bleu) RESET. 55
Figure 3.8 Comparaison de la distribution obtenue pendant le cyclage avec l'écriture en courant (mêmes données que sur la Figure 3.7 exceptés les 2 000 premiers cycles) et d'un exemple de distribution obtenue par écriture en tension (extrait de [12]) 56
Figure 3.9 Comparaison entre (au-dessus) le nuage de points de la Figure 3.7 et (en-dessous) le même nuage de points donc les points marqués virtuels par l'algorithme de vérification ont été retirés. La fenêtre de lecture ciblée par l'agorithme est 25 µA. L'agorithme n'est pas activé pendant les 2 000 premiers cycles correspondants au « seasoning ». 56
Figure 3.10 Comparaison entre les distributions de la Figure 3.8 et les distributions correspondantes après avoir appliqué l'algorithme de vérification pour différentes fenêtres de lecture (RM) cibles 10 µA, 15 µA, 20 µA et 25 µA. La distribution obtenue par écriture en tension (extrait de [12]), ainsi que sa distribution après vérification sont également rapportées pour comparaison. 56
Figure 3.11 Histogrammes représentant le nombre de cycles ayant nécessité un certain nombre de cycles de vérification pour différentes marges de lecture cibles. 56
Figure 4.1 Interface de la matrice mémoire 16k et cellule 1T1R. 66
Figure 4.2 Représentation de l'architecture de la source de courant intégrée et de l'ordre d'utilisation des sources unitaires pour différents exemples de courant de sortie. 67
Figure 4.3 Schéma de principe de la plateforme de test du circuit *MEMORIIC*. 68
Figure 4.4 De gauche à droite, schéma de principe d'un circuit de lecture, schéma d'une méthode de mesure de courant typique et schéma du circuit de lecture utilisé et intégré sur la carte mère. 69

Figure 4.5 Représentation des différents composants logiques, de leur attribution à l'ordinateur ou au micro-contrôleur et de leur communication. L'ordinateur et le micro-contrôleur fonctionnent de façon asynchrone. 71
Figure 4.6 Capture d'écran de l'interface du logiciel développé. On peut voir à gauche l'historique des opérations et les contrôles permettant de paramétrer et lancer une nouvelle opération, sur fond blanc une représentation graphique des opérations sélectionnées dans l'historique, en bas à droite une fenêtre permettant de voir l'intégralité de l'historique d'une cellule. 73
Figure 4.7 Répartition des huit groupes de cellules numérotées de 0 à 7. Chaque case mesure 16 cellules de côté et contient donc 128 cellules. 73
Figure 4.8 Conductances des éléments mémoires OxRAM juste après la formation de leur filament. 76
Figure 4.9 Conductances des cellules formées après 1 000 cycles. Deux distributions sont dessinées pour chaque expérience. Celle de gauche correspond à l'état de plus faible conductance après un RESET, celle de droite correspond au SET. 78
Figure 4.10 Tracé alternatif des données de la Figure 4.9. Ici les points des deux distributions sont appairés par leur adresse d'origine. Un score proportionnel à la différence de conductance entre les deux membres de la paire leur est associé. Leur ordonnée est leur score normal et leur abscisse leur conductance. Les deux points de chaque paire sont dessinés, l'un en bleu (HRS) et l'autre en rouge (LRS). 78
Figure 4.11 Stratégies de SET utilisées 81
Figure 4.12 Distributions cumulées des courants de lecture après 2 000 cycles. Les distributions HRS (gauche) et LRS (droite) sont toutes deux représentées. Les points de couleur transparente, respectivement pleine, correspondent au courant mesuré immédiatement après l'écriture, respectivement après une période de relaxation de 10 s. Les encarts correspondent aux histogrammes de la proportion d'éléments mémoires OxRAM ayant effectivement subi l'impulsion en abscisse (le complément ayant déjà atteint le seuil de vérification) pour les méthodes utilisant la vérification. 81
Figure 4.13 Représentation alternative, similaire à la Figure 4.10, des données de Figure 4.12. Contrairement à la Figure 4.10, la valeur de conductance HRS la plus élevée de chaque adresse sur les 1 000 derniers cycles est appairée avec la conductance LRS qui l'a précédée. 83
Figure 4.14 Histogramme de la population de cellules dans une certaine tranche de taux d'erreur en fonction de la méthode d'écriture utilisée (en haut), et histogramme I_R estimé comme optimum et utilisé pour le calcul du taux d'erreur (en bas). 85
Figure 4.15 Cartographie de la matrice. La nuance indique la stratégie utilisée dans la région, et la luminosité indique le taux d'erreur. 85
Figure 5.1 Illustration de l'évolution de la tension sur l'élément mémoire OxRAM et du courant dans l'élément mémoire OxRAM pour une écriture avec une source de tension et pour une écriture avec une source capacitive. 90
Figure 5.2 Illustration de l'évolution de la tension sur l'élément mémoire OxRAM et du courant dans l'élément mémoire OxRAM pour une écriture avec une source capacitive statique et pour une écriture avec une source capacitive rechargée. 92
Figure 5.3 Dessin des masques du circuit *CAPA-WR*. 95
Figure 5.4 Tension et courant dans l'élément mémoire OxRAM dans 3 simulations d'écriture. De gauche à droite, avec une source de tension et en RESET, avec une source capacitive, avec une source capacitive et le système de rechargement de la capacité. Dans ces simulations, l'élément mémoire OxRAM est représenté par un élément résistive de 50 kΩ en HRS changeant brusquement de résistance à un instant prédéterminé pour atteindre 5 kΩ en

LRS. La capacité utilisée est composée de 120 x 0,6 pF MOM et 5 x 3,8 pF MOS pour un total de 91 pF. 96
Figure 6.1 Arrangement avec un des deux pôles conducteurs mis en commun entre toutes les cellules. Il y a $N + 1$ connections externes avec N le nombre d'éléments mémoires OxRAM. 102
Figure 6.2 Arrangement avec deux groupes au sein desquels l'un des deux pôles conducteurs est mis en commun, et où le pôle restant est connecté avec l'homologue de l'autre groupe. Il y a $\frac{N}{2} + 2$ connections externes avec N le nombre d'éléments mémoires OxRAM. 102
Figure 6.3 Arrangement sous forme de matrice carrée. Les N cellules sont réparties sur un tableau de $\sqrt{N}$ lignes et $\sqrt{N}$ colonnes. Il y a $2\sqrt{N}$ connexions externes avec N le nombre d'éléments mémoires OxRAM. 102
Figure 6.4 Règle du 1/2 appliquée à la matrice. La cellule sélectionnée est la seule à voir une tension V. Les $2\sqrt{N}$ cellules partageant la ligne ou la colonne de cette dernière voient une tension de $\frac{V}{2}$. Les autres $\left(\sqrt{N}-1\right)^2$ cellules voient une tension nulle. Cette règle minimise le nombre de cellules non sélectionnées voyant une tension non nulle. 103
Figure 6.5 Règle du 1/3 appliquée à la matrice. La cellule sélectionnée est la seule à voir une tension V. Les $2\sqrt{N}$ cellules partageant la ligne ou la colonne de cette dernière voient une tension de $\frac{V}{3}$. Les autres $\left(\sqrt{N}-1\right)^2$ cellules voient une tension de $-\frac{V}{3}$. Cette règle minimise la valeur absolue de la tension parasite apparaissant sur les cellules non sélectionnées. 103
Figure 6.6 Exemple de matrice 1T1R. Seule la cellule sélectionnée (en noir) ne peut conduire du courant. 106
Figure 7.1 Comparaison entre une cellule 1T1R vue dans les précédents chapitres et une cellule 1TnR. Pour la cellule 1TnR, n éléments mémoires OxRAM sont connectés au même transistor, et n BL sont disponibles. 109
Figure 7.2 Illustration des différences entre GL$\perp$ (gauche) et SL$\perp$ (droite). Les lignes sont colorées selon leur potentiel, rouge: V, vert: $\frac{V}{2}$ et bleu: 0 V. Ces choix de potentiel correspondent à une stratégie d'écriture des éléments mémoires OxRAM représentés avec un fond rouge. Les flèches rouges en trait plein représentent les courants utiles à l'écriture et les traits en pointillés les courants de fuite. 110
Figure 7.3 Quatre stratégies de SET optimisant chacune un paramètre différent. Les éléments mémoires OxRAM que l'on essaie d'écrire ont un fond rouge. Les couleurs des lignes correspondent à leur potentiel, 0 V (bleu), V_L (vert), $2V_L$ (rouge), $3V_L$ (jaune). Les flèches rouges représentent les courants utiles (en trait plein) et parasites (en pointillé). 113
Figure 7.4 Représentation permettant la détermination des potentiels V_X^s et V_X^u. La zone grise représente leur lieu. L'axe des abscisses correspond au potentiel. Les potentiels ont été choisis selon les quatre stratégies de la Figure 7.3, et en appliquant (0.1) et (0.2) lorsque nécessaire en considérant $V_L = 0{,}7$ V, que les éléments mémoires OxRAM ont une conductance comprise entre 0 µS et 200 µS, que le transistor a une conductance constante de 2000 µS et que $n = 8$. 116
Figure 7.5 Cellule 1TnR sélectionnée (le sélecteur est passant) vue comme un réseau de résistance. Le courant i peut être observé à travers la SL. 117
Figure 7.6 À gauche, une représentation graphique de la contrainte $h < l^2$ en fonction des largeurs des intervalles de conductances associés aux états LRS et HRS. À droite, représentation des intervalles des lectures pour quatre exemples de choix de l et h particuliers avec, en haut, les états vus par une lecture absolue, en bas, les états vus par une lecture relative. Dans les exemples 1 et 2, la lecture relative est possible. Dans l'exemple 3 seule la lecture

absolue est possible. Dans l'exemple 4 les états HRS et LRS ne sont pas séparés et ne permettent aucune lecture. 118
Figure 7.7 Dessin des masques correspondants aux éléments mémoires OxRAM et aux niveaux métalliques les encadrant. Le trait en pointillé indique la limite du motif qui peut être répété par aboutement. 123
Figure 7.8 Dessin d'un transistor à trois doigts au sein d'une matrice de transistors de taille proche du minimum et respectant la contrainte de continuité des zones actives. L_G est la longueur de la grille, L_D la longueur inter-grille, W_G la largeur de la grille et W_e la largeur entre deux bandes de transistors. Le rectangle en pointillé dénote la maille associée à un transistor dans ce réseau carré de transistors. 124
Figure 7.9 Illustration de la problématique de dessin de deux transistors sur une même bande d'actif. A gauche, le cas où les deux transistors partagent leur source (cela pourrait également être leur drain). A droite, le cas où les deux transistors ne partagent ni leur source ni leur drain et requièrent donc une grille de séparation, noté B, polarisée pour ne pas laisser passer de courant entre les deux zones de diffusion correspondant à S_1 et S_2. 125
Figure 7.10 Dessin simplifié des niveaux transistors. Seul l'actif et les grilles sont dessinés. Le texte indique à quoi chaque zone de diffusion est logiquement connectée, avec S_x la source numéro x et X_x le nœud interne de la cellule numéro x. Le trait en pointillé indique l'étendue du motif qui peut être répété par aboutement après réflexion. 126
Figure 7.11 Dessin des niveaux d'interconnexion entre le front-end (les grilles, drains et sources des sélecteurs) et le niveau métallique M4 connecté aux éléments mémoires OxRAM (représenté sur la Figure 7.7). Les trois niveaux d'interconnexion sont représentés séparément en trait noir. Les traits gris représentent le niveau en-dessous de celui dessiné. Le trait en pointillé indique l'étendue du motif qui peut être répétée par aboutement après réflexion. 126
Figure 7.12 Vue d'ensemble du dessin du circuit *OX1TnR*. 128
Figure 7.13 Structure d'une matrice 1T0R. Chaque point représente une adresse. Les points bleus correspondent à un court-circuit en place d'un élément mémoire OxRAM. Ceux en gris indiquent l'absence de connexion. Les zones rouge pâle et verte en périphérie représentent les multiplexeurs d'accès aux BL en haut et en bas, et aux SL à gauche et à droite. Les multiplexeurs verts sont les paires court-circuitées entre elles. Les lignes et colonnes en gris foncé sont celles sur lesquelles les multiplexeurs d'accès sont court-circuités et donc celles sur lesquelles il ne doit y avoir aucune autre connexion. Enfin, les lignes en pointillé positionnées toutes les 8 colonnes séparent la matrice en 16 zones représentant chacune une colonne de cellule mémoire partageant une même GL. 130
Figure B.1 Chronogramme de l'écriture dans la valeur « Valeur à écrire » dans le registre à l'adresse « Adresse du registre ». 153
Figure B.2 Chronogramme d'une écriture en tension. 156
Figure B.3 Chronogramme d'une lecture. 156
Figure B.4 Chronogramme d'une écriture capacitive. 157
Figure C.1 Cellule 1TnR sélectionnée (le sélecteur est passant) vue comme un réseau de résistance. Le courant i peut être observé à travers la SL. 158
Figure C.2 Représentation des lieux de V_0 pour $h = 0.2$ et $l = 0.6$ pour, de gauche à droite, $n = 2$, 3 et 4. Le lieu correspondant à l'état composé uniquement de H n'est pas représenté car il recouvrirait l'intégralité du simplexe. 159

Liste des tableaux

Tableau 1.1 Exemple des caractéristiques typiques des différents niveaux mémoires. 15
Tableau 4.1 Description des opérations utilisées sur la plateforme de test du circuit *MEMORIIC*. . . 74
Tableau 4.2 Nomenclature des expériences de formation de filament sur le circuit *MEMORIIC* 76
Tableau 4.3 Conditions de cyclage 78
Tableau 4.4 Paramètres utilisés pour les stratégies de SET 81
Tableau 6.1 Tableau comparatif de différentes technologies mémoire 100
Tableau 7.1 Possibilités d'orientation des connexions au sein de la matrice 110
Tableau 7.2 Comparaison entre les architectures GL⊥ et SL⊥ 110
Tableau 7.3 Caractéristiques des matrices disponibles dans le circuit *OX1TnR* 128
Tableau A.1 Tableau exhaustif des instructions disponibles sur le micro-contrôleur de la plateforme de tests du circuit *MEMORIIC*. 150
Tableau B.1 Descriptif des plots d'accès du circuit *CAPA-WR* 152
Tableau B.2 Descriptif des registres du circuit *CAPA-WR* 153
Tableau B.3 Descriptif des entrées des registres du circuit *CAPA-WR* 153

Chapitre I : Introduction

I.1 La mémoire

De tout temps, la mémoire a été au cœur de la vie de l'Homme. La mémoire orale en est la forme la plus naturelle, innée chez l'Homme. Néanmoins, c'est en la dépassant que des civilisations naquirent. Au fil des âges, les supports de l'information ont pris diverses formes. La gravure sur pierre, parfois taillée, est la plus connue car elle a traversé les âges et est d'ailleurs toujours utilisée aujourd'hui sur les sépultures avec l'objectif de rendre l'inscription éternelle. D'autres supports ont été utilisés pour leurs avantages propres en regard de leur époque. Certains supports étaient choisis pour leur abondance comme l'argile, la pierre, l'os, le bambou, le bois, les ostraca[1] ou le papyrus puis le parchemin et le papier. Ces trois derniers présentent également l'avantage d'être facilement transportables, légers et denses en informations favorisant ainsi leur diffusion. D'autres supports tels que les tablettes de cire et la tablette blanche présentaient l'avantage d'être effaçables et réutilisables. Certains supports étaient choisis pour leur rareté ou leur valeur symbolique, par exemple les métaux, l'ivoire ou la soie. Ces supports ont évolué avec l'histoire et le progrès technique.

La révolution numérique n'a pas changé ce constat. Les différents supports de l'information ont divers avantages, mais aucun support de l'information ne les réunit tous. Ainsi, les types de support sont multiples et spécialisés. Quand on parle de stockage de l'information, les principales propriétés d'intérêt sont la densité volumique d'information ainsi que leur accessibilité. Cet intérêt n'est pas nouveau et s'illustre historiquement par le passage du volumen (rouleau de papyrus) au codex (cahier de parchemin) puis au livre, augmentant considérablement la densité d'information stockable par une bibliothèque tout en la rendant plus accessible grâce au système de page. De nos jours, le champion en terme de densité est le disque dur. A titre comparatif, la densité d'information d'un livre est de l'ordre[2] de 10 kb/cm³, alors que celle d'un disque dur est de l'ordre[3] de 100 Gb/cm³.

Cependant, les disques durs disparaissent petit à petit de notre quotidien au profit des SSD (*Solid-State*

[1] Un ostracon est un tesson de poterie ou un éclat de calcaire utilisé comme support d'écriture. Ces tessons ou éclats étaient sinon destinés au rebut, ce qui les rendait abondants.

[2] En considérant un livre de 500 pages avec environ 2 000 caractères par page dans une langue utilisant un alphabet codable sur 8 bits, cela donne 8 Mb par livre. L'épaisseur du livre est d'environ 3 cm sur des pages d'environ 15 cm par 21 cm pour un volume d'environ 1000 cm³. Donc la densité d'information est dans cet exemple de 8 kb/cm³, 5 kb/cm³ si l'on considère que chaque caractère de la langue ne porte que 5 bits d'information.

[3] En considérant un disque dur 2.5 pouces de 1 To (8 Tb), son volume est d'environ 70 cm³ (10 x 7 x 1 cm). Sa densité d'information est donc dans cet exemple d'environ 114 Gb/cm³.

Drive). La raison est que la densité[4] n'est pas l'unique propriété d'intérêt. Dans ce cas précis, on peut mettre en avant deux aspects, la rapidité et la résistance mécanique. N'ayant aucune pièce mobile, le SSD est plus robuste que le disque dur, il est donc plus pertinent dans des appareils portables qui représentent une part de plus en plus importante de notre environnement. La rapidité du SSD face au disque dur en fait également un support de choix lorsque l'usage requiert un accès fréquent aux données.

Cependant, la mémoire n'est pas qu'un support de l'information. Elle est également un des éléments constitutifs de toute machine à calculer telle qu'un ordinateur. La vitesse de la mémoire utilisée limite la vitesse d'exécution de la machine l'utilisant. Ainsi la rapidité d'une mémoire est également une propriété importante. Cette caractéristique se décline en deux grandeurs importantes, le débit qui correspond à une quantité d'informations transmise par unité de temps, et la latence qui correspond à la durée entre le début et la fin d'une opération.

L'exemple du disque dur et du SSD illustre le compromis entre rapidité d'accès et densité aujourd'hui au cœur des technologies mémoires. Une hiérarchie des mémoires s'est créée autour de ce compromis. Chaque application exploite un niveau précis de cette hiérarchie si bien qu'elle se retrouve dans l'architecture de nos systèmes.

Les mémoires les plus rapides sont utilisées par les unités de calculs. Leur densité, et *a fortiori* leur quantité disponible, est limitée. Pour permettre la réalisation de calcul requérant plus de mémoire que celle disponible, il est alors nécessaire d'utiliser une mémoire plus dense. Comme la réduction de vitesse n'est pas souhaitable, il est plus intéressant d'utiliser un mélange de technologies mémoires car aucunes n'est universellement meilleures.

Sur un certain intervalle de temps, seule une partie de l'information est utile à l'unité de calcul. Ainsi le reste de l'information peut être temporairement stocké dans une mémoire plus lente sans pénaliser la vitesse d'exécution à condition que, les données soient rapatriées à temps vers la mémoire rapide lorsque l'unité de calcul s'apprête à s'en servir et que le débit vu par la strate supérieure ne dépasse pas le débit disponible. Cette division peut être répétée pour former plusieurs niveaux de mémoire. Chaque niveau devient de plus en plus dense, mais de moins en moins rapide (en terme de latence et de débit) en s'éloignant du cœur de calcul. Cette division respecte la hiérarchie mémoire en adaptant la technologie mémoire utilisée aux caractéristiques du niveau. Les différents niveaux intermédiaires sont généralement qualifiés de mémoires tampons.

Pour illustrer ce principe, on peut regarder du côté des livres. Les bibliothèques ont une capacité limitée en particulier pour les rayonnages les plus accessibles. Les documents plus rarement consultés sont rangés de manière moins accessible et nécessitent une échelle et l'intervention d'un employé pour y accéder. Les documents encore plus anciens, ou encore moins susceptibles d'être lus sont rangés dans des archives. Dans ce cas on constate également que la quantité d'effort et d'énergie à dépenser pour accéder aux documents augmente à chaque niveau. Pour la mémoire numérique, on retrouve le même principe, où l'énergie requise croît avec la distance entre le cœur de calcul et le dispositif mémoire.

Ainsi on peut classer les technologies mémoires sur une même échelle mesurant leur place au sein du compromis densité/vitesse et leur place dans nos systèmes. En partant du cœur de calcul d'une machine grand public :

1. Les registres, implémentés en SRAM (*Static Random Access Memory*), stockent les données en cours de traitement ou les résultats des précédentes opérations.
2. Les différents niveaux de mémoires tampon (L1, L2 ou L3), généralement implémentés en SRAM,

[4]Techniquement, la quantité de données stockées dans un SSD par unité de volume est plus grande que celle d'un disque dur. Cependant, la dimension de ces appareils n'est généralement pas l'élément limitant, son prix l'est bien souvent plus. Or, la quantité de données stockées par un SSD par euro/dollar est plus faible que celle d'un disque dur.

servent de stockage court terme pour les opérations nécessitant plus de mémoire que les registres ne peuvent en stocker. Ils forment également un pont entre les niveaux qui l'entourent en essayant d'anticiper le besoin en récupérant une copie des parties de la mémoire vive externe prochainement utile. L'objectif est de cacher la latence de la mémoire vive externe. Leur différence de vitesse ne provient pas de l'utilisation d'une technologie de mémoire différente, mais de leur architecture système qui subit également le même compromis. Un autre facteur est leur distance au nœud de calcul, car l'espace proche de nœud est limité. Il est donc nécessaire de s'en éloigner pour avoir une grande surface disponible.

3. La mémoire vive externe (communément appelée RAM), implémentée en DRAM (*Dynamic Random Access Memory*), permet d'étendre la capacité mémoire de la puce intégrant le cœur en conservant une vitesse d'accès importante. Étant externe, l'espace disponible n'est plus un problème, au-delà du coût. Par contre, le délai introduit par l'interconnexion entre cette mémoire et le nœud de calcul le devient. C'est pour cela que leur architecture est principalement focalisée sur le débit plutôt que sur la latence.
4. Le stockage, aujourd'hui principalement implémenté par des SSD, première mémoire non volatile de cette liste, permet de stocker les données sur le moyen terme sans risquer d'en perdre en cas de perte de l'alimentation électrique ou d'extinction du système.
5. L'archivage, implémenté par des disques durs, permet le stockage de données mais avec une vitesse d'accès plus lente.

La Tableau 1.1 donne les ordres de grandeurs des caractéristiques de chaque niveau.

Dans cet empilement, les deux éléments centraux sont les mémoires tampons et la mémoire vive externe. Leurs caractéristiques sont aujourd'hui un des points limitant l'augmentation de la performance de nos machines.

Dans le modèle de mémoire utilisé aujourd'hui, l'espace mémoire est homogène. Du point de vue de la machine de calcul, toutes les adresses mémoires de la mémoire vive sont disponibles et équivalentes. La relation avec leur localisation physique est abstraite. La partie logicielle s'occupe de définir la façon dont les données doivent être transformées, et la partie matérielle est la seule responsable de leur organisation physique. L'avantage est qu'un même programme fonctionne aussi bien sur une machine ou sur une autre machine même si les caractéristiques du matériel varient fortement. Mais pour réussir cette abstraction, des systèmes complexes pénalisant généralement la latence des accès mémoires sont nécessaires. De plus, l'abstraction n'est jamais parfaite. En pratique, la localisation physique de données et leur trajet au sein du système ont des conséquences non triviales observables par le logiciel.

Tableau 1.1 – Exemple des caractéristiques typiques des différents niveaux mémoires.

	Capacité	Débit (Go/s)	Latence (cycles)
Registres[α]	~ 16 ko	–	–
Tampon L1[α]	32 ko	330	4
Tampon L2[α]	1 Mo	182	14
Tampon L3[α]	32 Mo	150	~ 50
DRAM[β]	~ 32 Go	57	~ 500
SSD	~ 2 To	1 – 10[γ]	1000 – 30000[δ]
HDD	~ 8 To	~ 0.2	~ 30000

[α] Pour un AMD Ryzen 9 7950X (architecture Zen 4) mesuré par [1].
[β] Pour une DRAM 128-bit DDR5-6 000 mesuré par [1].
[γ] Varie selon le type de connectivité entre le SSD et le reste du système.
[δ] Varie selon le modèle et selon la charge de travail. Ordre de grandeur issu de [2]

Ce problème est croissant. En effet, alors que le nombre d'opérations effectuables par seconde dans un cas idéal par une unique puce croît en moyenne de 50 % par an [3], le débit de la mémoire externe ne croît que de 23 % par an [3]. Ainsi, si toutes les données traitées doivent être extraites de la mémoire vive externe, il est aujourd'hui nécessaire d'effectuer plus de 100 opérations par mot mémoire pour maintenir le maximum d'opérations réalisables par seconde par la puce [3]. Les mémoires tampons permettent de réduire cette contrainte lorsque les données peuvent être localement réutilisées.

La latence de la mémoire vive, c'est à dire le délai entre la décision d'aller chercher la donnée dans la mémoire et le moment où elle devient effectivement disponible stagne, voire augmente jusqu'à 4 % par an[5] [3]. La latence n'est pas un problème dans l'absolu puisque les mémoires tampons sont supposées la rendre invisible au yeux de la machine de calcul. C'est d'ailleurs pour cette raison qu'elle a tendance à ne pas évoluer, voire à régresser, car elle est mise de côté au profit du débit mémoire. Cependant, la machine n'est pas omnisciente, et déterminer suffisamment en avance quelle partie de la mémoire sera requise dans le futur plus ou moins proche n'est pas une tâche facile. Pour donner un ordre de grandeur, entre le début de la requête d'accès à des données stockées dans la mémoire vive externe à leur arrivée au niveau du cœur de calcul, une puce peut aujourd'hui réaliser environ 10 000 opérations [3]. Cela signifie également qu'un accès mémoire déclenché trop tardivement, car imprévisible par exemple, peut pénaliser très fortement la performance. De nombreux systèmes sont en place pour limiter l'impact d'une telle occurrence, mais leurs capacités sont limitées et un code particulièrement conçu pour les provoquer peut virtuellement faire régresser nos machines d'environ 23 ans[6] d'évolution. C'est un exemple des conséquences concrètes de l'imperfection de l'abstraction du système mémoire physique. Aujourd'hui, la performance d'un logiciel dépend en premier lieu de sa capacité à utiliser ce système mémoire à son avantage, ou plutôt à sa capacité à ne pas le mettre en difficulté.

Pour l'instant, ce problème a été repoussé par une ingéniosité toujours plus grande trouvant des solutions pour perfectionner cette abstraction et rendre les cas de dégradation importante des performances de plus en plus rares. Cela a cependant un coût en terme de complexité des systèmes, et donc en terme de coût de développement. Cela rend leur analyse plus complexe et augmente la probabilité de présence d'un bug exploitable à des fins malveillantes[7]. Une solution proposée par [3] est d'abattre la frontière à l'origine du modèle mémoire est de rendre le logiciel responsable de son utilisation de la mémoire physique, en lui donnant les outils nécessaires. Cette solution a beaucoup de sens dans le cas du calcul de hautes performances, puisque l'intégralité d'une machine est généralement dédiée à un unique code. Cependant, dans le cas général où la machine doit partager ses ressources entre plusieurs logiciels indépendants, son application en pratique semble complexe.

Une solution alternative consiste à essayer de trouver de nouvelles technologies mémoire, aux propriétés différentes, ouvrant la porte à de nouvelles architectures mémoire. C'est dans cette idée que la recherche autour des mémoires résistives s'inscrit.

I.2 Mémoire résistive

Les mémoires discutées jusqu'à maintenant, à l'exception des disques durs, ont un point en commun, elles représentent et stockent l'information par la présence ou l'absence de charge électrique. Les SRAM utilisent des circuits CMOS bistables chargeant une de leurs branches et déchargeant l'autre. Les DRAM stockent des charges dans des capacités enterrées dans le substrat. Les mémoires Flash stockent des charges dans un élément métallique entièrement entouré d'isolant. Les charges y sont

[5] L'augmentation de la latence est principalement due à l'ajout de niveau de mémoire tampon et à la complexification de système mémoire en général, requérant de plus en plus d'intelligence et multipliant les lieux de prise de décision.

[6] 23 années d'évolution à +50 % d'opération par seconde par an correspond à une multiplication du nombre d'opérations par seconde par 11222.

[7] Exemples: Spectre [4], Meltdown [5], TLBleed [6], Cachebleed [7]. Exemple et citation extraits de [8].

introduites ou extraites par effet tunnel. Ainsi, ces mémoires forment une sorte de famille que l'on pourrait qualifier de mémoires basées sur la charge électrique. Les disques durs stockent quant à eux leur information dans la texture magnétique de leurs disques.

Les mémoires résistives constituent une autre famille de mémoire. Comme son nom l'indique, ces membres utilisent l'état de résistivité, ou de conductivité selon le point de vue, d'un matériau comme vecteur d'information. L'information peut être lue par application d'une tension aux bornes du matériau et l'observation du courant circulant dedans. De plus, l'état de résistivité du matériau doit pouvoir être modifié pour permettre l'écriture, et stable pour plusieurs valeurs pour permettre la rétention de l'information. Les différentes combinaisons de matériaux et de phénomènes physiques répondant à ces contraintes forment les différents types de mémoire résistive.

On peut distinguer trois grandes catégories de mémoire résistive (ReRAM[8]) :

1. **Les mémoires magnétiques (MRAM[9])**
   Elles sont composées d'un empilement de couches magnétiques dont deux séparées par une couche de matériau isolant. L'une des couches magnétiques autour de l'isolant a son état magnétique, son aimantation, fixé par l'empilement alors que celui de la seconde couche est libre. Par effet de magnétorésistance géante, la résistivité de l'empilement dépend fortement de l'orientation relative de l'aimantation des deux couches encadrant l'isolant. La direction de l'aimantation de la couche libre peut être modifiée par application d'un champ magnétique, via la circulation d'un courant dans une ligne proche par exemple. Elle peut également être modifiée par le courant traversant la jonction elle-même par le phénomène de transfert de spin, c'est la spécificité des STT-MRAM (*Spin-Tranfert Torque MRAM*). Dans le cadre d'une mémoire dense, cette seconde technique d'écriture limite le risque d'écriture des dispositifs adjacents.
2. **Les mémoires à changement de phase (PCM[10])**
   Elles sont composées d'une couche d'un matériau ayant plusieurs phases stables de résistivité différentes, tels que les verres de chalcogénure, prise entre deux électrodes métalliques. L'information est stockée dans la phase du matériau observable via leur différence de résistivité. L'écriture se réalise par effet Joule permettant le chauffage du matériau et ainsi son changement de phase.
3. **Les mémoires oxyde (OxRAM[11])**
   Elles sont composées d'une couche d'un oxyde, généralement métallique, prise entre deux électrodes métalliques. Par application d'un champ électrique suffisamment fort sur la structure, des ions oxygènes peuvent être générés par effet électrochimique puis déplacés laissant derrière eux une succession de lacunes d'oxygène favorisant la conduction au travers de l'oxyde. Avec une bonne maîtrise du courant maximum, cet état est partiellement réversible par application d'un champ électrique opposé favorisant la recombinaison des lacunes et des ions oxygène. De manière similaire, d'autres matériaux peuvent être utilisés pour former par réaction électrochimique des chemins conducteurs dans un matériau. C'est le cas de la CBRAM par exemple[12].

Cette thèse se focalise sur les mémoires OxRAM, plus particulièrement à base d'oxyde d'hafnium (HfO_2) dont la structure et le fonctionnement seront plus amplement décrits dans le Chapitre II.

I.3 Cas d'application

Dans cette partie, et la fin de ce chapitre, nous nous intéressons aux cas d'application des mémoires résistives discutés dans la littérature. Cette évaluation des domaines d'utilisation a pour objectif

[8] ReRAM pour *Resistive Random Access Memory*
[9] MRAM pour *Magnetic Random Access Memory*
[10] PCM pour *Phase Change Memory*
[11] OxRAM pour *Oxide Random Access Memory.*
[12] CBRAM pour *Conductive Bridge Random Access Memory*

d'établir un contexte applicatif pour la suite de cette thèse. En particulier, cela permettra de nous informer sur l'importance de chacune des caractéristiques de la mémoire du point de vue applicatif.

I.3.1 Mémoire embarquée

Une force de l'OxRAM HfO_2 est sa forte intégrabilité au processus de fabrication CMOS. Elle est donc particulièrement intéressante en tant que mémoire embarquée. Sa persistance lui permet de replacer la mémoire flash embarquée qui souffre d'un besoin de forte tension lors de l'écriture (environ 10 V). L'OxRAM est donc préférable pour les nœuds les plus avancés. De nombreuses démonstrations de circuit ont été faites [9] [10] [11] [12] [13].

Comme l'empilement OxRAM peut être réalisé entre les niveaux de métaux, une perspective est d'utiliser l'OxRAM en complément des mémoires SRAM pour rendre le stockage persistant en cas de perte de la source d'énergie, tout en conservant la rapidité de la SRAM [14].

Les caractéristiques de l'OxRAM pourraient la rendre utilisable en tant que mémoire vive persistante qui contrairement à l'actuel DRAM ne nécessiterait pas de rafraîchissement continu. Cependant, l'endurance de l'OxRAM est encore un peu faible pour ce cas d'application. L'avance de développement, ainsi que les progrès continus de la DRAM ne laissent pour l'instant aucune ouverture au développement de l'OxRAM, ou à d'autres types de mémoire, en tant que mémoire vive persistante.

I.3.2 Accélération matérielle

L'OxRAM, et plus généralement les mémoires résistives sont plébiscitées dans de nouvelles architectures de calcul. On les dénomme souvent architectures « non von-Neumann » en opposition aux architectures « von-Neumann » utilisées dans la quasi-intégralité de nos ordinateurs. La différence entre ces architectures est la relation entre la mémoire de la machine et ses unités de calcul. L'architecture « von-Neumann » sépare la mémoire des unités de calcul et les connecte par un bus. Cette architecture est très flexible, cependant, avec l'amélioration de l'électronique numérique, l'énergie consommée par une opération de calcul ne cesse de diminuer alors que celle utilisée pour transporter les données diminue beaucoup moins vite et devient dominante. Ainsi, aujourd'hui, la réduction de la puissance consommée doit passer par une amélioration de ces interconnexions. Ce problème ne concerne pas uniquement l'efficacité énergétique, il concerne également la performance. En effet, l'élément limitant la conception de système numérique de haute performance est la consommation électrique et la capacité à le refroidir [15]. L'émergence des *system on package* qui consiste à intégrer plusieurs circuits intégrés au sein d'un même boîtier est une des réponses apportées par l'industrie à cette problématique d'interconnexion consommant trop d'énergie.

L'idée des architectures dites « non von-Neumann » est de distribuer la mémoire au sein des unités de calcul et d'ainsi réduire la quantité de données à transférer. En ce sens, elle n'est pas très différente de l'objectif des mémoires caches insérées aux différents niveaux des architectures « von-Neumann ». Comme elles, elles espèrent conserver les données qui seront réutilisées plus tard près de l'unité de calcul pour éviter leur transfert à répétition qui prendrait du temps et de l'énergie. Mais, quoi que l'on fasse, il y aura toujours un compromis entre proximité et capacité. Plus la quantité de données à faire interagir ensemble est grande, plus il faudra investir de l'énergie dans son transport, architecture von-Neumann ou pas.

Ainsi, le point clé n'est pas vraiment la classe d'architecture en elle-même, mais plus exactement le taux de réutilisation des données atteint au plus proche des éléments de calcul. Ce taux dépend énormément de l'application et de l'implémentation du calcul sur l'architecture. Les principales innovations sont donc sur la spécialisation des architectures (et de leur programmation) à différentes tâches ainsi que sur la gestion de la distributions des données et des résultats intermédiaires lorsque le problème requiert

une grande interaction entre les données[13]. C'est pourquoi ces architectures sont plutôt des circuits d'accélération que de nouveaux processeurs généraux. Ces problématiques sont en effet exactement celles rencontrées par les concepteurs des processeurs graphiques [15].

Cette question du taux de réutilisation est d'autant plus importante lorsque l'on souhaite comparer une architecture à base de SRAM (généralement von-Neumann) à celles basées sur des mémoires résistives (généralement non-von-Neumann). En effet, la consommation d'énergie d'une écriture d'une mémoire résistive [16] est au moins 2 ordres de grandeur supérieure à celle d'une SRAM [17]. Ainsi, pour un faible taux de réutilisation des données, une architecture à base de SRAM a toutes les chances d'être plus énergétiquement performante qu'une architecture utilisant des mémoires résistives. Par contre, lorsque la réutilisation des données est importante, par exemple dans la limite où l'écriture des données n'est réalisée qu'une seule fois, les mémoires résistives pourront être compétitives.

a) Memristors et logique

Le memristor est le quatrième composant passif élémentaire prédit par L. Chua [18], et il est approximativement implémenté par les différents types de mémoires résistives. Des unités de calcul logique d'un nouveau genre utilisant des « memristors » ont été proposées [19] [20] [21] [22] [23]. Ces unités de calcul sont composées de memristors qui, dans l'idée, servent à la fois de mémoire et d'unité de calcul logique. L'idée est d'effectuer l'opération logique en écrivant des memristors au travers d'autres.

Donnons un exemple : pour effectuer une opération non-ou (NOR) un memristor de sortie peut être pré-programmé à un niveau de conductance élevé, un 1 logique. Puis plusieurs memristors pré-programmés à un niveau logique représentant l'entrée de l'opération NOR peuvent être connectés en parallèle entre eux et en série au memristor de sortie. En appliquant une tension entre les deux pôles libres, une tension dépendante de l'entrée de l'opération apparaît au point milieu. Lorsque tous les memristors d'entrée sont au 0 logique représenté par une forte résistivité, la tension chute essentiellement sur ces memristors, et le memristor de sortie ne voit qu'une faible tension. Au contraire, lorsqu'au moins un des memristors d'entrée est à l'état logique 1 de faible résistivité, une plus forte tension apparaît sur le memristor de sortie. En choisissant l'orientation des memristors correctement, on peut donc faire en sorte que le memristor de sortie soit écrit à l'état correspondant à l'opération NOR sur les entrées en appliquant simplement une impulsion électrique [19]. L'implémentation avec des dispositifs réels pose cependant quelques complications [23]. Nous relevons trois limites à cette approche :

1. **Quantité de copies nécessaire**

   Les memristors d'entrée et de sortie doivent pouvoir être connectés d'une manière précise pour effectuer l'opération. Ainsi l'opération ne peut pas être effectuée directement sur des memristors arbitrairement choisis parmi l'ensemble de la mémoire. Il est donc fort probable qu'en pratique, un nombre important d'opérations de copie ou de déplacement de données soit nécessaire. Cela diminue l'intérêt de l'architecture et la rend finalement peu différente d'une architecture von-Neumann. Ce problème est d'autant plus important que la taille des assemblages de memristors est limitée, réduisant d'autant plus l'ensemble des opérations effectuables sans copie préliminaire. Notons que le positionnement des données pourrait éventuellement être optimisé par un compi-

[13] Considérons deux problèmes différents, mais de taille similaire. Dans un premier cas, nous souhaitons multiplier une matrice carrée de taille N avec elle-même, dans le second cas, nous souhaitons sommer N^2 valeurs. Dans les deux cas, la quantité de données totale est N^2. Mais si l'on considère le nombre de données avec lequel chacune devra être mise en relation, on constate une large différence. Dans le cas de la somme, cette quantité est 1, la valeur de l'accumulateur. Dans le cas de la matrice, cette quantité est environ N. Cette différence prend de l'importance lorsque l'on souhaite découper le problème en morceaux pour pouvoir le faire rentrer dans une mémoire de taille limitée. Dans le cas de la somme, chaque morceau est indépendant. Mais dans le cas de la matrice, chaque morceau devra être au moins assez grand pour accueillir les N valeurs à faire interagir avec chaque valeur, sans quoi une certaine quantité de données devra être dupliquée, et donc transférée à plusieurs reprises depuis une mémoire plus grande (soit au travers du temps, soit de l'espace).

lateur pour minimiser ces copies. Toutefois, cela reste spéculatif et ce devrait être un des premiers éléments à vérifier pour quiconque souhaite s'investir dans cette approche.

2. **Quantité d'énergie nécessaire**
   Les comparaisons effectuées par [22] montrent que le seul résultat expérimental considéré a une consommation 4 ordres de grandeur supérieure à l'opération utilisant un circuit CMOS. Certes, la consommation du circuit CMOS n'inclut pas la lecture des données ni leur éventuelle écriture. Mais on pourrait en dire autant du cas des memristors qui, d'après le premier point, ont peu de chance de pouvoir réaliser l'opération souhaitée sans copie de données. De plus, les perspectives ne sont pas bonnes. D'après ces mêmes comparaisons [22], les résultats simulés, représentant probablement un cas idéal, sont au moins 2 ordres de grandeur supérieurs en consommation. Ce n'est pas très étonnant, en effet, le calcul avec les memristors repose indirectement sur leur écriture. Or l'écriture des memristors est plus consommatrice qu'une opération CMOS (de calcul, ou de transport de données), car en plus de devoir charger les capacités parasites, un courant statique important doit être dépensé.

3. **Durée de l'opération de calcul**
   Le délai de l'opération est au moins 3 ordres de grandeur supérieur à l'opération CMOS [22]. Ce délai provient du temps d'écriture des memristors. Bien que celui-ci peut, en théorie, toujours être accéléré en augmentant la tension d'écriture [24], en pratique, la limite de tension des nœuds plus avancés et le risque de perturber les memristors d'entrée ou à proximité limitent ce délai d'écriture. Ce délai est particulièrement gênant car pendant toute l'opération, aucune autre opération ne peut être effectuée sur le même ensemble de memristors (ceci incluant tous les memristors partageant des connections avec ceux en opération). Comme la plupart des opérations sont composites et nécessitent une séquence d'opération logique simple, les délais s'accumulent. Comme pour le premier point, quelqu'un pourrait argumenter que ces problèmes peuvent être réduits par une bonne organisation de la mémoire et un assez bon compilateur, mais cela reste hypothétique.

b) Opération logique à la lecture

La lecture des mémoires résistives s'effectue en mesurant le courant passant dans la cellule lorsqu'une faible tension lui est appliquée. Dans la plupart des cas d'utilisation, il n'est pas nécessaire de mesurer la valeur de la résistivité de la cellule, mais uniquement de la comparer à une référence pour en obtenir l'état logique. Que se passerait-il si on essayait de lire plusieurs dispositifs à la fois, laissant leur courant s'ajouter ?

Dans le cas de deux éléments mémoires OxRAM connectés en parallèle, chacun avec deux états possibles, trois états pourraient être observés. En utilisant deux références, on pourrait distinguer ces trois états. Le premier correspondant à l'état 00, le second à 01 ou 10 (confondus entre eux), et le dernier à 11. Chacune des références seule serait suffisante pour nous donner le résultat d'une opération logique différente. La référence distinguant 00 des autres états nous informe sur le OU logique entre les deux entrées, alors que la référence distinguant 11 nous donne le résultat du ET logique [25] [26] [27].

Cette idée peut être étendue à un nombre plus élevé d'éléments mémoires OxRAM, ou à des éléments programmés sur plus de 2 états. La marge de lecture est alors d'autant plus réduite que l'on doit considérer deux niveaux de résistivité distincts. La première référence donne toujours le résultat du OU logique alors que la dernière donne le ET logique.

De la même manière, il est également possible de calculer la somme numérique des valeurs mémorisées [27]. Cependant, pour pouvoir sommer des valeurs sur plus de 1 bit, il est nécessaire d'utiliser des éléments mémoires OxRAM multi-niveaux avec une répartition linaire des niveaux de conductance. L'utilisation de plusieurs tensions d'excitation pour pondérer les mémoires et représenter des valeurs multi-bits peut également être envisagée en remplacement du multi-niveaux. Cependant, la dynamique

de la tension de lecture est limitée par le bas par la précision de la comparaison des courants, et par le haut par la tension seuil d'écriture de la mémoire.

Ces idées souffrent d'inconvénients similaires à la logique basée sur l'écriture de memristors. Cependant, elles ont l'avantage de ne pas requérir l'écriture des dispositifs pour effectuer l'opération. Elles ne souffrent donc pas des limites de consommation énergétique ni d'un délai dû à l'écriture des memristors. Néanmoins, cela signifie également que le résultat n'est pas mémorisé et donc indisponible pour effectuer une opération subséquente sans être écrite au préalable, ce qui présentera à nouveau ces inconvénients.

c) Multiplication-accumulation analogique

En poussant l'idée précédente un peu plus loin, on peut réaliser une opération particulièrement importante aujourd'hui, une multiplication-accumulation (MAC pour « multiply-accumulate »). Cette opération est au cœur du produit scalaire, de la multiplication matricielle ou encore de l'évaluation de polynôme. Elle est très utilisée, particulièrement dans les applications graphiques, de traitement du signal ou d'intelligence artificielle.

Comme dans l'idée précédente, l'objectif est d'utiliser les lois d'Ohm[14] et de Kirchhoff pour effectuer l'opération de multiplication et d'addition respectivement. La tension appliquée sur l'élément mémoire OxRAM est convertie en courant par multiplication par la conductance de l'élément mémoire OxRAM, puis les courants des cellules en parallèle s'ajoutent. Le courant résultant peut être mesuré pour obtenir le résultat de l'opération. Du fait d'une une organisation de matrice d'élément mémoire OxRAM courante, la crossbar, cette même opération peut aisément être effectuée en parallèle pour différentes valeurs de conductance pré-écrites dans les éléments mémoires OxRAM. Cela permet d'effectuer une opération de multiplication vecteur-matrice sur une durée similaire à une seule multiplication-accumulation [28].

L'opération au sein de la mémoire est en elle-même énergétiquement efficace, ce qui constitue son attrait principal [28]. Cependant, des DAC et ADC sont nécessaires pour imposer les tensions d'entrée et mesurer les courants de sortie. Ils permettent de faire le lien entre le monde numérique et cette opération analogique. Ils créent un compromis entre la résolution de l'opération et leur consommation. En pratique ils réduisent fortement l'efficacité énergétique du système. En effet, leur consommation domine la consommation due au courant circulant dans les cellules [29] [30]. Ils occupent également une surface importante. Dans le cas où la sortie d'une de ces opérations est utilisée comme entrée d'une autre, il est possible de se passer de cette conversion analogique-numérique. Une solution proposée est d'utiliser des signaux en modulation de largeur d'impulsion (PWM pour *Pulse Width Modulation*) pour communiquer analogiquement entre deux opérations successives [31].

A contrario des opérations de logique présentées dans la sous-partie précédente, ici l'opération est analogique. La précision de la conductance de l'élément mémoire OxRAM est donc un facteur important [32] [33]. Cela représente un défi pour la procédure d'écriture, et particulièrement pour l'OxRAM qui présente une faible prédictibilité [34].

Cette technique de calcul est très appréciée pour les applications d'intelligence artificielle. En effet, les réseaux neuronaux sont résilients aux erreurs [30] [35]. Cependant, des dégradations de performance importantes par rapport aux implémentations logicielles peuvent être observées [36]. Une solution est d'entraîner le réseau pour plus de résilience en incorporant un modèle du bruit présent dans la version matérielle.

L'entraînement peut également être réalisé sur la version matérielle. Cela assurerait une bonne adaptation au bruit provenant du circuit. Cependant, la mise-à-jour des poids du réseau peut poser

[14]Dans l'état de faible résistance, le mode de conduction de l'élément mémoire OxRAM est Ohmique.

problème [37] [38]. En effet, les algorithmes d'entraînement reposent sur la linéarité et la symétrie de l'évolution des poids. Mais la caractéristique d'évolution de la conductance des OxRAMs n'est ni linéaire ni symétrique. De plus, lors de l'évaluation, un faible nombre de niveaux de conductance (forte quantification des poids) peut suffire à obtenir une inférence de qualité (par exemple dans un réseau neuronal binaire [39]). Mais un nombre de niveaux de conductance important (au moins 64 [37]) est nécessaire lors de l'étape d'accumulation des modifications des poids pendant l'entraînement.

Une alternative est d'utiliser des algorithmes d'entraînement particulier permettant une approche hybride entre entraînement logiciel et entraînement matériel. Par exemple, un algorithme de meta-apprentissage tel que présenté par [40] peut être utilisé pour pré-entraîner le réseau dans un état où seules quelques itérations d'entraînement sur la tâche et le matériel précis permettent d'atteindre des performances similaires à un réseau entièrement entraîné dans ces conditions [41]. Cela permet donc de réduire considérablement la quantité d'entraînement nécessaire à faire sur la version matérielle, tout en gardant les avantages que cela apporte en terme de compensation des non-idéalités mais également en terme de protection des données.

Cette méthode est performante pour l'évaluation des réseaux, en particulier lorsque l'intégralité des poids peut être stockée dans le module de calcul. Cela élimine toute nécessité de déplacer ces informations. Cela représente un réel avantage face au système von-Neumann. Cependant, l'entraînement directement au sein du circuit représente encore un défi. Le taux de réutilisation des données s'effondre lors de l'entraînement puisque les nouveaux poids doivent être réécrits ou modifiés à chaque étape.

Les réseaux de neurones sont la principale application étudiée dans la littérature car leur utilisation massive de l'opération de multiplication-accumulation et leur résilience aux erreurs en font des candidats idéaux. Mais d'autres applications sont envisageables. L'élément généralement bloquant est la précision atteignable. Cette précision peut faire défaut à la fois au sein de l'élément mémoire OxRAM, mais également au niveau de l'ADC lorsqu'un grand nombre de produits est additionné. Une solution proposée est de faire essentiellement du numérique en analogique [42] [43]. Pour ce faire, on quantifie les niveaux de résistivité et les signaux d'entrée (réduisant ainsi les contraintes sur les DAC) et on adapte la résolution de l'ADC pour tout juste couvrir la dynamique de la multiplication-accumulation. En pratique, l'ADC peut également être l'élément limitant, on effectue alors le choix des niveaux de quantification pour satisfaire les contraintes de l'ADC. On obtient ainsi un système plus robuste mais avec une dynamique ou une résolution faible (pas plus de quelques bits pour les poids par exemple [44]).

Pour augmenter la dynamique de travail, on peut décomposer une valeur (d'entrée ou de poids) dans une base adaptée, et considérer chaque chiffre dans cette base comme un élément sur lequel le calcul est fait. On peut ensuite décomposer l'opération sur les éléments de cette nouvelle représentation, les effectuer, et recomposer le résultat. Un plus grand nombre d'opérations doit être effectué, mais cette méthode permet d'atteindre la haute précision (au moins 16 bits expérimentalement et 64 bits en simulation d'après [42]) nécessaire à la plupart des applications en-dehors de l'intelligence artificielle telle que la résolution d'équations différentielles [42].

I.4 Conclusion

Nous avons vu que les mémoires sont au cœur de nos machines de calculs et représentent aujourd'hui une contrainte pour l'amélioration de leur performance. C'est pour cette raison que de nouveaux dispositifs mémoires, aux combinaisons de propriétés nouvelles, sont recherchés et étudiés dans l'objectif d'ouvrir la voie à de nouvelles architectures.

L'un des éléments identifié comme bloquant est la forte ségrégation entre mémoire et unité de calcul des architectures dites « von-Neumann ». La non-volatilité et la co-intégrabilité des mémoires résis-

tives sont un atout pour permettre le développement d'architectures intégrant la mémoire et les unités de calculs de manière plus proche et efficace.

La représentation de l'information via la résistivité variable des dispositifs de mémoire résistive permet le calcul d'une opération de multiplication-accumulation dans le domaine analogique par la programmation d'un des vecteurs de l'opération dans la mémoire. Dans le cas ou ce vecteur est fixe, ou changeant peu fréquemment, l'opération peut être conduite sur un grand nombre de vecteurs de manière efficace. Comme cette opération est très utilisée dans les calculs matriciels à la base de nombreux codes de calcul, ou d'inférence de réseau de neurones, elle justifie le développement de circuits d'accélération spécialisés. Un des besoins identifié pour cette application est la capacité de programmer la conductivité des mémoires de manière continue, ou du moins sur un certain nombre de niveaux, et avec peu de dispersion.

L'OxRAM présente naturellement une forte dispersion de sa conductance. L'objectif de la première partie de cette thèse (Chapitre II – Chapitre V) est donc d'étudier différentes manières d'écrire l'OxRAM dans l'objectif d'en améliorer les caractéristiques, en particulier en terme de maîtrise de la dispersion des états.

Le Chapitre II introduit le fonctionnement d'une OxRAM, définit ses caractéristiques importantes et présente les méthodes d'écriture décrites par la littérature.

Les Chapitre III et Chapitre IV présentent des résultats expérimentaux d'une méthode d'écriture utilisant comme excitation une source de courant. Le Chapitre III discute de l'aspect théorique de cette approche et donne des résultats préliminaires obtenus sur un circuit issu d'une précédente thèse n'ayant pas été conçu pour ces expériences. Le Chapitre IV présente quant à lui la conception d'un circuit de test spécifique à cette technique d'écriture et les résultats qu'il a permis d'obtenir.

Le Chapitre V s'intéresse à l'utilisation d'une décharge capacitive comme excitation pour l'écriture des OxRAM. Un circuit spécifique est conçu et décrit pour l'expérimentation de cette technique dans un environnement intégré. Ce circuit ayant été conçu peu de temps avant la fin de cette thèse, il n'a pas pu être testé. Ce chapitre ne présente donc pas de résultat expérimental.

La seconde partie (Chapitre VI – Chapitre VII) de cette thèse est consacrée à la densité, une des caractéristiques fondamentales de toute mémoire.

Le Chapitre VI constitue l'introduction de cette partie. Il explique les problématiques liées à l'augmentation de la densité d'intégration des OxRAM. Il présente également les principales stratégies utilisées pour construire un agrégat dense d'éléments mémoires OxRAM.

Le Chapitre VII se focalise sur un genre d'architecture dit 1TnR, un hybride des deux principales stratégies utilisées pour agréger les éléments mémoires OxRAM visant à augmenter la densité. Une nouvelle architecture 1TnR ayant fait l'objet d'un dépôt de demande de brevet est décrite. Puis son implémentation dans une technologie CMOS FDSOI 22 nm est discutée, particulièrement du point de vue de son implémentation physique au travers du dessin des masques de fabrication.

Bibliographie

[1] C. Lam, « AMD's Zen 4, Part 2: Memory Subsystem and Conclusion ». [En ligne]. Disponible sur: https://chipsandcheese.com/p/amds-zen-4-part-2-memory-subsystem-and-conclusion?utm_source=publication-search

[2] I. Jo, « Toward Ultra-Low Latency SSDs: Analyzing the Impact on Data-Intensive Workloads », *Electronics*, vol. 13, nº 1, 2024, doi: 10.3390/electronics13010174.

[3] McCalpin et D. John, « Memory bandwidth and System Balance in HPC Systems », in *International Conference for High Performance Computing, Networking, Storage, and Analysis*, 2016.

[4] P. Kocher *et al.*, « Spectre Attacks: Exploiting Speculative Execution ». [En ligne]. Disponible sur: https://arxiv.org/abs/1801.01203

[5] M. Lipp *et al.*, « Meltdown ». [En ligne]. Disponible sur: https://arxiv.org/abs/1801.01207

[6] B. Gras, K. Razavi, H. Bos, et C. Giuffrida, « Translation Leak-aside Buffer: Defeating Cache Side-channel Protections with TLB Attacks », in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, MD: USENIX Association, août 2018, p. 955-972. [En ligne]. Disponible sur: https://www.usenix.org/conference/usenixsecurity18/presentation/gras

[7] Y. Yarom, D. Genkin, et N. Heninger, « CacheBleed: a timing attack on OpenSSL constant-time RSA », *Journal of Cryptographic Engineering*, vol. 7, p. , 2017, doi: 10.1007/s13389-017-0152-y.

[8] G. Dessouky *et al.*, « HardFails: Insights into Software-Exploitable Hardware Bugs », in *28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA: USENIX Association, août 2019, p. 213-230. [En ligne]. Disponible sur: https://www.usenix.org/conference/usenixsecurity19/presentation/dessouky

[9] P. Jain *et al.*, « 13.2 A 3,6 Mb 10,1 Mb/mm2 Embedded Non-Volatile ReRAM Macro in 22 nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0,5 V with Sensing Time of 5 ns at 0,7 V », in *2 019 IEEE International Solid- State Circuits Conference - (ISSCC)*, févr. 2019, p. 212-214. doi: 10.1109/ISSCC.2019.8662393.

[10] J. Yang *et al.*, « 24.2 A 14 nm-FinFET 1 Mb Embedded 1T1R RRAM with a 0.022µm2 Cell Size Using Self-Adaptive Delayed Termination and Multi-Cell Reference », in *2 021 IEEE International Solid- State Circuits Conference (ISSCC)*, févr. 2021, p. 336-338. doi: 10.1109/ISSCC42613.2021.9365945.

[11] C.-C. Chou *et al.*, « An N40 256K×44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance », in *2 018 IEEE International Solid - State Circuits Conference - (ISSCC)*, févr. 2018, p. 478-480. doi: 10.1109/ISSCC.2018.8310392.

[12] B. Giraud *et al.*, « Benefits of Design Assist Techniques on Performances and Reliability of a RRAM Macro », in *2 023 IEEE International Memory Workshop (IMW)*, mai 2023, p. 1-4. doi: 10.1109/IMW56887.2023.10145984.

[13] C. Ho *et al.*, « Integrated HfO2-RRAM to achieve highly reliable, greener, faster, cost-effective, and scaled devices », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 2. doi: 10.1109/IEDM.2017.8268314.

[14] Y. Liu *et al.*, « 4.7 A 65 nm ReRAM-enabled nonvolatile processor with 6× reduction in restore time and 4× higher clock frequency using adaptive data retention and self-write-termination nonvolatile logic », in *2 016 IEEE International Solid-State Circuits Conference (ISSCC)*, janv. 2016, p. 84-86. doi: 10.1109/ISSCC.2016.7417918.

[15] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, et D. Glasco, « GPUs and the Future of Parallel Computing », *IEEE Micro*, vol. 31, nº 5, p. 7-17, sept. 2011, doi: 10.1109/MM.2011.89.

[16] G. Sassine, C. Cagli, J.-F. Nodin, G. Molas, et E. Nowak, « Novel Computing Method for Short Programming Time and Low Energy Consumption in HfO2 Based RRAM Arrays », *IEEE Journal of the Electron Devices Society*, vol. 6, p. 696-702, 2018, doi: 10.1109/JEDS.2018.2830999.

[17] C.-C. Wang, R. G. B. Sangalang, I.-T. Tseng, Y.-J. Chiu, Y.-C. Lin, et O. L. J. A. Jose, « A 1.0 fJ energy/bit single-ended 1 kb 6T SRAM implemented using 40 nm CMOS process », *IET Circuits, Devices & Systems*, vol. 17, nº 2, p. 75-87, 2023, doi: 10.1049/cds2.12141.

[18] L. Chua, « Memristor-The missing circuit element », *IEEE Transactions on Circuit Theory*, vol. 18, nº 5, p. 507-519, sept. 1971, doi: 10.1109/TCT.1971.1083337.

[19] S. Kvatinsky *et al.*, « MAGIC—Memristor-Aided Logic », *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, nº 11, p. 895-899, nov. 2014, doi: 10.1109/TCSII.2014.2357292.

[20] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, et R. S. Williams, « 'Memristive' switches enable 'stateful' logic operations via material implication », *Nature*, vol. 464, nº 7290, p. 873-876, avr. 2010, doi: 10.1038/nature08940.

[21] S. Kvatinsky, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, et U. C. Weiser, « Memristor-Based Material Implication (IMPLY) Logic: Design Principles and Methodologies », *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, nº 10, p. 2054-2066, oct. 2014, doi: 10.1109/TVLSI.2013.2282132.

[22] T. Zanotti, F. M. Puglisi, et P. Pavan, « Smart Logic-in-Memory Architecture for Low-Power Non-Von Neumann Computing », *IEEE Journal of the Electron Devices Society*, vol. 8, p. 757-764, 2020, doi: 10.1109/JEDS.2020.2987402.

[23] S. M. Laube et N. TaheriNejad, « Device Variability Analysis for Memristive Material Implication ». Consulté le: 24 octobre 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2101.07231

[24] L. Chua, « Five non-volatile memristor enigmas solved », *Applied Physics A*, vol. 124, nº 8, p. 563, juill. 2018, doi: 10.1007/s00339-018-1971-0.

[25] W.-H. Chen *et al.*, « A 16 Mb dual-mode ReRAM macro with sub-14 ns computing-in-memory and memory functions enabled by self-write termination scheme », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 28. doi: 10.1109/IEDM.2017.8268468.

[26] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, et Y. Xie, « Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories », in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, juin 2016, p. 1-6. doi: 10.1145/2897937.2898064.

[27] E. Esmanhotto *et al.*, « Experimental demonstration of Single-Level and Multi-Level-Cell RRAM-based In-Memory Computing with up to 16 parallel operations », in *2 022 IEEE International Reliability Physics Symposium (IRPS)*, mars 2022, p. P8–1–P8–4. doi: 10.1109/IRPS48227.2022.9764474.

[28] J. Chen, J. Li, Y. Li, et X. Miao, « Multiply accumulate operations in memristor crossbar arrays for analog computing », *Journal of Semiconductors*, vol. 42, nº 1, p. 13104, janv. 2021, doi: 10.1088/1674-4926/42/1/013104.

[29] J. M. Correll *et al.*, « An 8-bit 20.7 TOPS/W Multi-Level Cell ReRAM-based Compute Engine », in *2 022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, juin 2022, p. 264-265. doi: 10.1109/VLSITechnologyandCir46769.2022.9830490.

[30] W. Wan *et al.*, « 33.1 A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models », in *2 020 IEEE International Solid- State Circuits Conference - (ISSCC)*, févr. 2020, p. 498-500. doi: 10.1109/ISSCC19947.2020.9062979.

[31] H. Jiang, W. Li, S. Huang, et S. Yu, « A 40 nm Analog-Input ADC-Free Compute-in-Memory RRAM Macro with Pulse-Width Modulation between Sub-arrays », in *2 022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, juin 2022, p. 266-267. doi: 10.1109/VLSITechnologyandCir46769.2022.9830211.

[32] A. Glukhov *et al.*, « Statistical model of program/verify algorithms in resistive-switching memories for in-memory neural network accelerators », in *2 022 IEEE International Reliability Physics Symposium (IRPS)*, mars 2022, p. 3. doi: 10.1109/IRPS48227.2022.9764497.

[33] V. Milo *et al.*, « Multilevel HfO2-based RRAM devices for low-power neuromorphic networks », *APL Materials*, vol. 7, nº 8, p. 81120, août 2019, doi: 10.1063/1.5108650.

[34] C.-X. Xue *et al.*, « A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices », *Nature Electronics*, vol. 4, nº 1, p. 81-90, janv. 2021, doi: 10.1038/s41928-020-00505-5.

[35] M. Hu *et al.*, « Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine », *Advanced Materials*, vol. 30, nº 9, p. 1705914, 2018, doi: 10.1002/adma.201705914.

[36] V. Milo *et al.*, « Accurate Program/Verify Schemes of Resistive Switching Memory (RRAM) for In-Memory Neural Network Circuits », *IEEE Transactions on Electron Devices*, vol. 68, nº 8, p. 3832-3837, août 2021, doi: 10.1109/TED.2021.3089995.

[37] S. Yu, « Neuro-inspired computing with emerging nonvolatile memorys », *Proceedings of the IEEE*, vol. 106, nº 2, p. 260-285, févr. 2018, doi: 10.1109/JPROC.2018.2790840.

[38] Y. Zhang *et al.*, « Optimized Programming Scheme Enabling Symmetric Conductance Modulation in HfO_2 Resistive Random-Access Memory (RRAM) for Neuromorphic Systems », *IEEE Electron Device Letters*, vol. 43, nº 8, p. 1203-1206, août 2022, doi: 10.1109/LED.2022.3182494.

[39] M. Bocquet *et al.*, « In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks », in *2 018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 20. doi: 10.1109/IEDM.2018.8614639.

[40] C. Finn, P. Abbeel, et S. Levine, « Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks », *CoRR*, 2017, [En ligne]. Disponible sur: http://arxiv.org/abs/1703.03400

[41] M. Pallo *et al.*, « On Chip Customized Learning on Resistive Memory Technology for Secure Edge AI », in *2 025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2025, p. 1-3. doi: 10.23919/VLSITechnologyandCir65189.2025.11074789.

[42] M. A. Zidan *et al.*, « A general memristor-based partial differential equation solver », *Nature Electronics*, vol. 1, nº 7, p. 411-420, juill. 2018, doi: 10.1038/s41928-018-0100-6.

[43] C.-X. Xue *et al.*, « 24.1 A 1 Mb Multibit ReRAM Computing-In-Memory Macro with 14,6 ns Parallel MAC Computing Time for CNN Based AI Edge Processors », in *2 019 IEEE International Solid-State Circuits Conference - (ISSCC)*, févr. 2019, p. 388-390. doi: 10.1109/ISSCC.2019.8662395.

[44] C. Li *et al.*, « Analogue signal and image processing with large memristor crossbars », *Nature Electronics*, vol. 1, nº 1, p. 52-59, janv. 2018, doi: 10.1038/s41928-017-0002-z.

Chapitre II : Écriture de l'Oxram

II.1 Introduction

Ce chapitre a pour objectif de présenter les aspects relatifs à l'écriture des OxRAM. Il permet d'introduire des notions utilisées dans les Chapitre III, Chapitre IV et Chapitre V.

Avant de passer en revue les techniques d'écriture utilisées dans la littérature, une description élémentaire de la physique de l'OxRAM est entreprise. Après une discussion de la composition physique ainsi que des processus physiques mis en jeu lors de l'écriture, la caractéristique quasi-statique d'une cellule OxRAM est présentée. Un ensemble de paramètres importants est défini et leurs relations décrites. L'importante question de la variabilité dans les OxRAM est ensuite discutée. Enfin, un ensemble de caractéristiques d'intérêt, à la fois à l'échelle d'une cellule OxRAM et à l'échelle d'un système de mémoire complet est défini.

Les méthodes d'écriture utilisées dans la littérature sont ensuite passées en revue, en commençant par celle utilisée en laboratoire pour obtenir les caractéristiques quasi-statiques présentées, jusqu'aux plus avancées, utilisées dans des systèmes mémoires complets ou des applications émergentes comme les réseaux de neurones. Enfin, nous discuterons de différents types de sources (au sens électrique du terme) utilisables pour l'écriture pour introduire les aspects particuliers qui seront le sujet des Chapitre III, Chapitre IV et Chapitre V.

II.2 Éléments constituants et considérations générales

II.2.1 Structure du dispositif et description physique élémentaire

Un élément mémoire OxRAM se présente sous la forme d'un empilement métal/oxyde/métal. La symétrie de l'empilement est cassée par d'un côté une couche métallique à forte affinité avec l'oxygène (Ti pour notre sujet d'étude), et une autre plus inerte (TiN ici). Une telle structure présente divers états de conductivité entre les deux électrodes métalliques. En particulier, à l'état initial, après la fabrication, peu de défauts sont présents dans l'oxyde qui est alors essentiellement isolant. Cependant il est possible d'augmenter grandement la conductivité de l'oxyde.

Un champ électrique est imposé dans l'oxyde à l'aide des électrodes métalliques. Des paires ion oxygène/trou d'oxygène sont alors générées. Le champ électrique provoque la dérive des ions oxygène. En choisissant bien l'orientation du champ, ces ions peuvent être conduits jusqu'à l'interface présentant une forte affinité avec l'oxygène. Ils sont alors absorbés [45]. Les trous d'oxygène restés piégés dans l'oxyde contribuent à l'augmentation de la conductivité du milieu en favorisant d'abord la conduction par effet Pool-Frenkel (conduction par émission de piège en piège, ici de trou en trou). L'augmentation

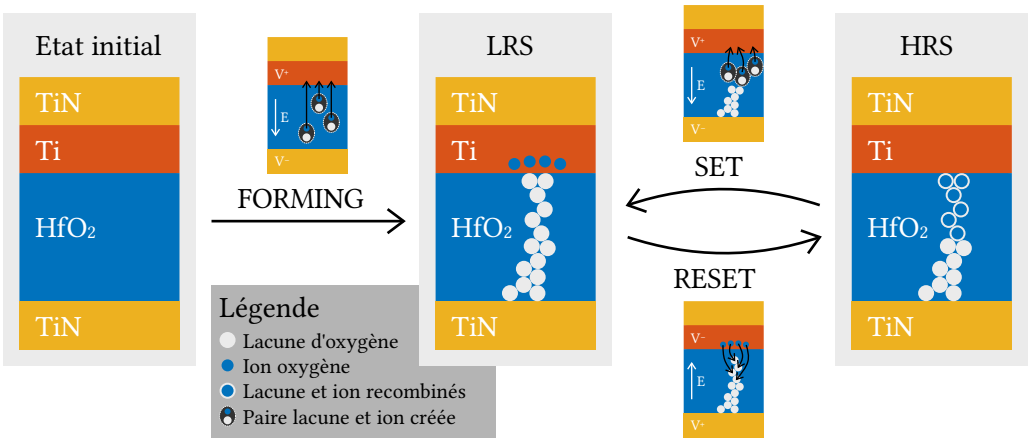


Figure 2.1 – Illustration des différents états et transitions entre ceux-ci dans un élément mémoire OxRAM.

de la conductivité d'une partie de l'épaisseur de l'oxyde tend à amplifier le champ électrique local dans l'autre partie (en considérant la différence de potentiel des électrodes constante). Cette amplification locale du champ électrique accélère la réaction localement. La formation des trous n'est donc pas homogène dans l'ensemble de l'oxyde, mais elle est localisée autour des sites de nucléation initiaux. Cette réaction conduit alors à la formation d'un filament de trous d'oxygène entre les deux électrodes [46] [47]. La conduction électrique entre les deux électrodes présente alors un comportement ohmique [48] et est beaucoup plus élevée qu'à l'état initial. Cet état est stable même à champ électrique nul.

La différence de potentiel est le facteur principal contrôlant la réaction d'écriture. La réaction tend à évoluer jusqu'à ce que la différence de potentiel atteigne une valeur « à l'équilibre » qui dépend de l'échelle de temps de l'observation [49]. Ainsi, si la différence de potentiel est directement fixée par une source de tension, la réaction continue jusqu'à la destruction totale de l'oxyde. La réaction doit donc être arrêtée avant en diminuant cette différence de potentiel. Généralement, un élément « limitant » le courant est ajouté en série de l'empilement pour modérer la réaction.

Si la réaction n'est pas trop avancée, elle est partiellement réversible. L'application d'un champ électrique opposé tend à dissoudre le filament en favorisant la recombinaison des ions oxygène stockés dans le métal avec les trous proches de l'interface. La conductivité dans l'oxyde est fortement diminuée et perd son caractère ohmique. Cependant le filament n'est pas intégralement dissolu, et la conductivité est plus forte qu'à l'état initial. Cela signifie également que la réaction de re-formation du filament nécessitera une différence de potentiel moins forte car elle profitera de l'amplification locale du champ électrique dès le début.

On distingue donc trois états :
1. L'état initial, ou neuf qui présente la plus faible conductivité
2. L'état écrit, ou LRS[15] qui présente la plus forte conductivité, de type ohmique
3. L'état effacé, ou HRS[16] qui présente une conductivité intermédiaire entre celles des deux autres états

Le cycle de vie d'un élément mémoire OxRAM est illustré sur la Figure 2.1. Il consiste à passer de l'état neuf à l'état écrit par une opération dite de formation (FORMING), d'écrit à effacé par une opération d'effacement (RESET), puis de nouveau à l'état écrit par une opération d'écriture (SET).

Différents matériaux d'oxyde métallique peuvent être utilisés pour la fabrication des éléments mémoires OxRAM. Le matériau considéré ici est l'oxyde d'hafnium. Ce matériau a l'avantage d'être

[15]LRS = Low Resistive State = État de faible résistance
[16]HRS = High Resistive State = État de haute résistance

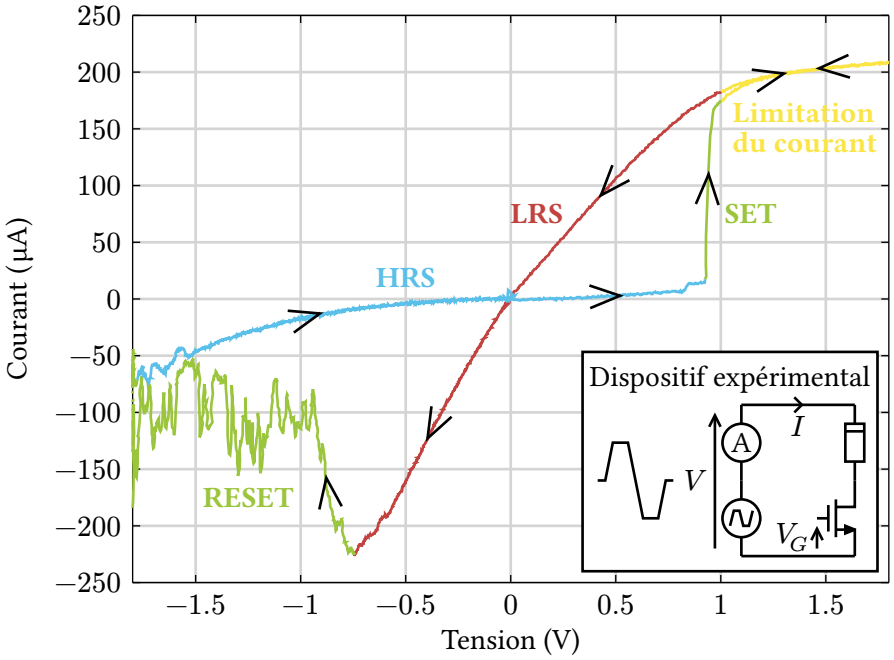


Figure 2.2 – Caractéristique quasi-statique d'une cellule 1T1R. Les flèches sur le cycle indiquent sa direction. Le temps de montée et descente de l'impulsion de tension est 350 kV/s (3,5 V en 10 µs) pour le SET et 180 kV/s (1,8 V en 10 µs) pour le RESET, avec un palier de 10 µs à la valeur maximale.

déjà utilisé dans le processus de fabrication des transistors CMOS en tant que matériau à forte constante diélectrique. C'est donc un matériau déjà bien étudié d'un point de vue du processus de fabrication, qui facilite l'intégration d'un processus de fabrication d'éléments mémoires OxRAM en co-intégration d'un nœud CMOS existant. De plus, l'oxyde d'hafnium riche en défauts présente de très bonnes caractéristiques pour le fonctionnement d'un élément mémoire OxRAM. C'est le matériau le plus utilisé aujourd'hui pour cette application. Une revue des différents matériaux considérés dans la littérature pour la fabrication des éléments mémoires OxRAM est disponible dans [47].

II.2.2 Cellule 1T1R, caractéristique quasi-statique et variabilités

La cellule OxRAM la plus répandue est la cellule dite 1T1R pour « un transistor une résistance » qui sont connectés en série. La « résistance » est en réalité l'élément mémoire OxRAM. Le « transistor » est quant à lui l'élément limitant le courant et permettant de modérer la réaction d'écriture. Le potentiel de la grille du transistor permet également de sélectionner, ou au contraire d'isoler, l'élément mémoire OxRAM parmi d'autres lorsqu'ils sont interconnectés sous forme de matrice mémoire.

La Figure 2.2 montre un exemple de caractéristique quasi-statique d'une cellule 1T1R. Pour l'obtenir, une impulsion triangulaire de tension quasi-statique est appliquée sur la cellule 1T1R, et le courant passant à travers elle est mesuré. Cette caractéristique est bipolaire, une tension positive permet d'écrire l'élément mémoire OxRAM de la cellule, alors qu'une tension négative permet de l'effacer. On peut noter quelques paramètres importants sur cette caractéristique :

- V_{SET} – la tension appliquée sur la cellule 1T1R lorsque le SET se produit;
- $V_{\mathrm{SET}}^{\mathrm{eff}}$ – la tension nécessaire pour que l'écriture se produise[17];
- V_{RESET} – la tension maximale utilisée lors de l'effacement;
- $V_{\mathrm{RESET}}^{\mathrm{eff}}$ – la tension où l'effacement commence;

[17] Ici, V_{SET} et $V_{\mathrm{SET}}^{\mathrm{eff}}$ correspond à la même tension, mais cela n'est pas nécessairement le cas, par exemple lors de l'utilisation d'impulsions carrées pour l'écriture.

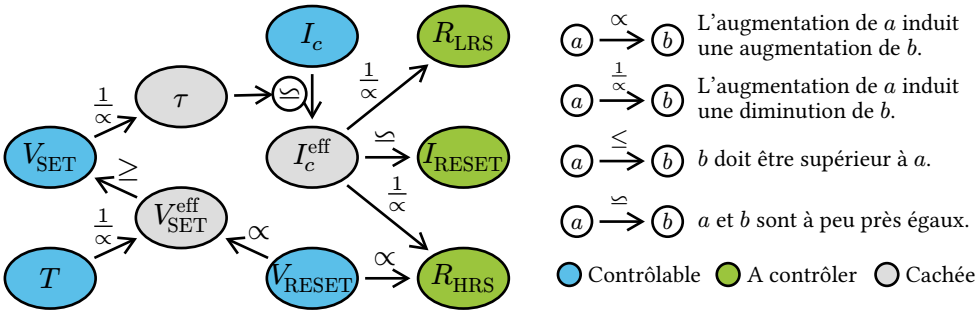


Figure 2.3 – Illustration des relations entre les paramètres identifiés. T est l'ordre de grandeur du temps d'écriture et τ est le temps caractéristique de la réaction d'écriture. Les grandeurs sur fond bleu sont contrôlables par l'expérimentateur. Celles sur fond vert sont celles que l'on cherche à contrôler. Celles sur fond gris ne sont ni contrôlables, ni aisément observables. La flèche partant de (τ) indique que τ influence la relation entre I_c et I_c^{eff}.

- I_c – le courant limite imposé (contrôlé par le transistor);
- I_c^{eff} – le courant limite atteint lors de l'écriture[18];
- I_{RESET} – le courant atteint lors de l'effacement;
- R_{LRS} / Γ_{LRS} – la résistance/conductance absolue perçue dans l'état écrit ou LRS;
- R_{HRS} / Γ_{HRS} – la résistance/conductance absolue perçue dans l'état effacé ou HRS (dépendant de la tension de lecture car non-ohmique).

La Figure 2.3 représente quant à elle les relations que l'on peut établir entre ces paramètres. En particulier, on a :

- $V_{\text{SET}}^{\text{eff}}$ augmente lorsque l'échelle de temps[19] considérée diminue [49]
- I_{RESET} est proportionnel à I_c^{eff}, et généralement égaux [50] [51]
- R_{LRS} est inversement proportionnel à I_c^{eff} [50] [52]
- R_{HRS} est de même inversement proportionnel à I_c^{eff} [50]
- R_{HRS} est également positivement relié à V_{RESET} [50] [52]
- $V_{\text{SET}}^{\text{eff}}$ augmente lorsque V_{RESET} augmente [52]
- R_{LRS} n'est pas directement influencé par V_{SET}, mais peut l'être indirectement[20] [51].

La Figure 2.4 présente une superposition d'un ensemble de caractéristiques quasi-statiques provenant de différentes cellules. Elle illustre la variabilité observée. On peut distinguer deux types de variabilité, celles provenant des différences entre chaque cellule en raison de dérive de fabrication par exemple, et celles intrinsèques à l'élément mémoire OxRAM que l'on constate généralement de cycle-à-cycle. Ainsi la Figure 2.4 montre le cumul des deux effets. On peut voir que la dispersion est assez élevée dans le cas de l'OxRAM.

La variabilité cycle-à-cycle est principalement due au caractère discret de la matière. Les particules composant le filament conducteur en LRS ou les pièges facilitant la conduction en HRS sont en nombre limité. Ainsi l'apparition d'une de plus ou une de moins, ou le déplacement de l'une d'elles peut avoir un effet observable important. Expérimentalement, on observe en effet que la variabilité cycle-à-cycle

[18] En théorie I_c^{eff} et I_c sont égaux, mais des effets parasites, comme la décharge rapide d'une capacité parasite peuvent entraîner une différence.

[19] L'échelle de temps dédiée à l'écriture. Si l'on souhaite que l'écriture se produise dans un temps limité de 100 ns, $V_{\text{SET}}^{\text{eff}}$ sera plus grand que si l'on souhaitait qu'elle se produise dans un intervalle de 1 s par exemple.

[20] V_{SET} peut indirectement influencer R_{LRS}. La dynamique d'écriture devient plus rapide lorsque V_{SET} augmente. Ainsi, pour V_{SET} suffisamment élevé, la dynamique d'écriture peut être plus rapide que la relaxation du couple RC entre la cellule 1T1R et la capacité parasite. Cela a pour conséquence d'avoir I_c^{eff} supérieur à I_c à cause du courant de décharge de la capacité qui vient s'ajouter au courant passant dans le transistor limitant le courant. Cela se traduit également par un I_{RESET} supérieur. [51]

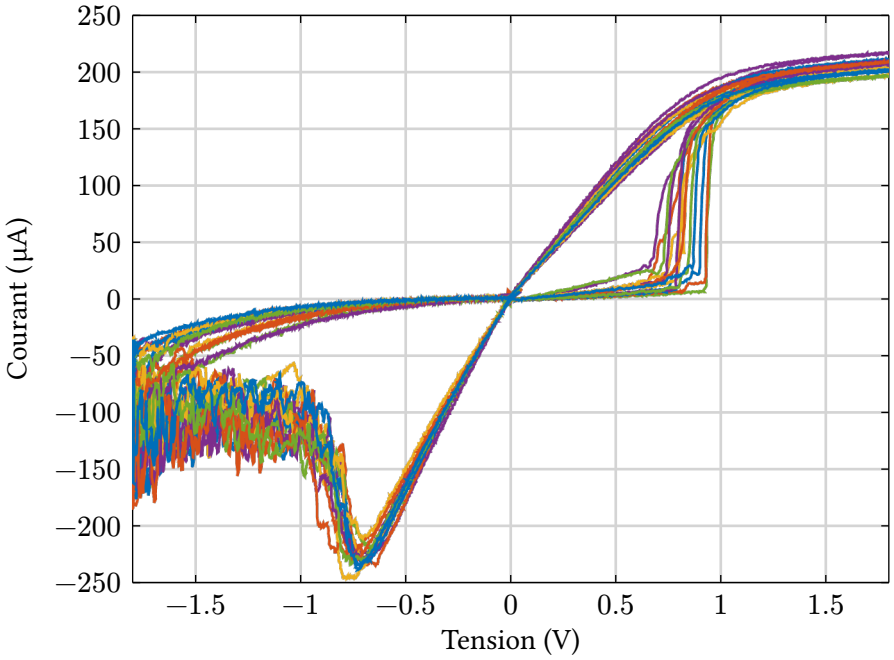


Figure 2.4 – Caractéristique quasi-statique de quinze cellules 1T1R. Les conditions d'écriture sont identiques pour l'ensemble des quinze cellules et identiques aux conditions décrites dans la Figure 2.2.

est d'autant plus forte que I_c^{eff} est faible [53] [54]. En effet, plus I_c^{eff} est faible, plus le filament est « fin », donc plus le nombre de particules le composant est faible, et donc plus l'influence de chacune sur l'observation est grande.

La variabilité cellule-à-cellule provient quant à elle de dérive de fabrication, en particulier des paramètres tels que l'épaisseur des films fins, surtout celui de l'oxyde, la stœchiométrie de l'oxyde, l'état de surface des interfaces, la quantité de défaut, etc.

En pratique, pour les OxRAM, la variabilité cycle-à-cycle domine la variabilité cellule-à-cellule si l'on ne considère que l'élément mémoire OxRAM. Toutefois, il est nécessaire de prendre également en compte les variations de caractéristique du transistor, en particulier lorsque celui-ci est utilisé comme limiteur de courant. La variabilité du transistor se traduit par une incertitude sur I_c^{eff} d'autant plus grande que le transistor est petit. La variabilité cellule-à-cellule peut alors devenir prépondérante en LRS [54]. On peut le constater sur la Figure 2.4, dans la zone où la tension est supérieure à 1,5 V, là où la caractéristique est principalement définie par le transistor (car il est en saturation).

Bien que l'état général (LRS ou HRS) de l'élément mémoire OxRAM soit stable, une variabilité temporelle est également présente. Encore une fois, la dimension discrète du problème est au cœur de la question. Ainsi plus I_c^{eff} est faible moins l'état est stable dans le temps. Il y a deux effets temporels particulièrement notables. D'un côté, il y a un bruit en créneaux (ou RTN pour Random Telegraph Noise) correspondant à l'écrantage causé par un électron aléatoirement et temporairement piégé dans un trou d'oxygène [55] [56] [57]. Il dépend fortement du diamètre du filament. D'un autre côté, il y a le phénomène de relaxation où la conductance du filament présente une instabilité temporaire juste après l'écriture. L'amplitude de variation de la conductance décroît rapidement avec le temps. Ce phénomène est attribué au fait que la conductance observée dépend à la fois du nombre de trous d'oxygène constituant le filament, mais aussi de leur disposition dans l'espace. Ainsi un même nombre de trous peut mener à des états de conductance très différents [58]. Juste après l'écriture, le système n'est pas encore nécessairement à l'équilibre. Avec le temps le système se relaxe jusqu'à atteindre son

équilibre. La nouvelle configuration, bien que constituée d'un nombre similaire de trous d'oxygène, peut alors présenter une conductance très différente de celle initiale.

II.2.3 Caractéristiques importantes d'une cellule 1T1R

Le cas général d'utilisation d'une OxRAM est celui d'une cellule mémoire. Dans un premier temps, nous ne considérons qu'une cellule seule. On distingue alors différentes caractéristiques importantes pour la cellule.

1. **Prédictibilité**
   La prédictibilité représente l'inverse de la variabilité cycle-à-cycle. Elle représente le niveau de précision avec lequel nous sommes capables de modifier le niveau de conductance de l'élément mémoire OxRAM.

2. **Endurance**
   L'endurance représente le nombre de cycles d'écriture et d'effacement qu'une cellule peut subir avant que ses caractéristiques ne se détériorent au point de ne plus permettre un bon fonctionnement du système l'utilisant [51]. L'importance de l'endurance dépend de l'application visée, plus particulièrement du nombre d'écritures par unité de temps et de la durée de vie du produit. Par exemple, pour une même durée de vie, l'utilisation en tant que mémoire vive (plusieurs écritures par seconde) demande une plus forte endurance que l'utilisation en tant que mémoire de stockage persistante (moins d'une écriture par jour).

3. **Rétention**
   La rétention représente la capacité de la cellule à conserver son état au fil du temps. L'OxRAM ne nécessite pas d'alimentation pour conserver son état. Néanmoins, il peut avoir tendance à dériver à cause de l'agitation thermique ou de diverses perturbations électriques, électro-magnétiques ou radiatives. De part sont fonctionnement, l'OxRAM est faiblement sensible aux perturbations électro-magnétiques ou radiatives. La rétention en LRS a été observée comme corrélée positivement au niveau de conductivité en LRS. [59].

4. **Densité d'état**
   La densité d'état représente la quantité d'informations stockable dans une cellule. Avec l'OxRAM, l'information est stockée sous forme de niveau de résistivité. La densité d'état représente le nombre de niveaux de résistivité distincts à laquelle la cellule 1T1R peut être programmée et lue. Un minimum de deux états est nécessaire et équivaut à un 1 bit d'information[21]. Un nombre plus important d'états permet d'augmenter la densité d'information stockable. La densité d'état est donc proportionnelle à la densité d'information et permet de réduire le coût par bit de la mémoire. A titre d'exemple, les SSD (« Solid-State Drives ») les moins chers utilisent aujourd'hui des cellules QLC (pour « Quad-Level Cell ») programmables dans 16 états différents, soit 4 bits d'information.
   On peut également envisager le stockage de l'information de manière analogique en l'encodant dans le niveau de résistivité. Il est alors nécessaire de s'intéresser au bruit généré par la variabilité de l'OxRAM.

Une cellule mémoire n'est généralement pas utilisée seule. On souhaite intégrer un grand nombre de cellules ensemble pour former une mémoire. On distingue alors des caractéristiques importantes pour cet ensemble.

1. **Uniformité**
   L'uniformité représente l'inverse de la variabilité dispositif-à-dispositif. Une forte uniformité permet l'utilisation de très nombreuses cellules au sein du même circuit. Au contraire, une faible

[21] Dans un système complet utilisant des méthodes de réduction d'erreur par exemple, le nombre de bit d'information par cellule, au niveau du système, peut être inférieur à 1. Cependant, au niveau de la cellule, un nombre minimum de deux niveaux distincts est nécessaire pour qu'elle soit utile.

uniformité impose d'avoir de plus grandes marges sur les autres critères pour compenser leurs variations et conserver un système fonctionnel sur toute la gamme de cellules. On peut noter qu'une calibration cellule par cellule n'est pas envisageable car cela demanderait une mémoire de calibration plus grande que celle engendrée par les cellules.

2. **Densité**
   La densité représente le nombre de cellules intégrables par unité de surface. Elle est donc proportionnelle à la densité d'information, et comme la densité d'état, elle est fortement liée au coût par bit de la mémoire. Elle influence également fortement la capacité mémoire envisageable. Dans une cellule 1T1R, le transistor requiert une certaine surface de silicium. L'élément mémoire OxRAM peut quant à lui être intégré entre deux niveaux de métal, au-dessus du transistor. Ainsi la densité est limitée par le plus grand élément constituant la cellule. Le transistor est généralement cet élément limitant [60].

3. **Taux d'erreur**
   Différents taux d'erreur peuvent être définis. Ici, nous considérons le taux d'erreur global égal à la proportion moyenne de lecture donnant une réponse ne correspondant pas à la valeur précédemment écrite. Ce taux inclut donc les erreurs provenant de toute la chaîne, aussi bien de l'écriture que de la lecture. Le taux d'erreur doit toujours être mis en relation avec le nombre de dispositifs mis en jeu au sein du système. En effet, un taux d'erreur de 10^{-6} erreur par bit assure un quasi-bon fonctionnement d'une cellule. Cependant lorsque plusieurs millions de cellules sont mises à contribution, le taux d'erreur du système augmente de façon significative, soit 1 erreur par megabit dans cet exemple.

4. **Fenêtre de lecture**
   La fenêtre de lecture représente la largeur de la plage de conductance libre ne correspondant à aucun état valide. Elle permet de séparer les plages de conductance correspondantes aux différents états et aussi de limiter le risque de confusion. Plus la fenêtre de lecture est grande, plus la marge d'erreur du circuit de lecture peut être grande. Cela permet par exemple de réaliser des lectures plus rapides. Comme la variabilité est une notion statistique et qu'une mémoire utilise un grand nombre de dispositifs (plusieurs millions), les cas rares doivent être considérés car il seront probablement réalisés. Ainsi, une fenêtre de lecture est souvent définie pour un taux d'erreur défini.

Comme remarqué par [54], la prédictibilité de l'élément mémoire OxRAM est limitée. Ainsi, une certaine plage de conductance doit être allouée à chaque état souhaité. Cela impose donc un compromis entre la densité d'état et la fenêtre de lecture. L'uniformité rentre aussi en jeu dans ce compromis.

La densité et le taux d'erreur ne sont *a priori* pas reliés. Cependant l'utilisation de **codes correcteurs d'erreurs** crée un compromis entre les deux. En effet, ces codes permettent la correction ou la détection d'une partie des erreurs en contrepartie d'une redondance dans les données stockées. La redondance vient réduire la densité d'information effective, alors que la possible correction vient diminuer le taux d'erreur. Ces codes sont communément utilisés à tous les niveaux des systèmes mémoires. Ils permettent d'assurer un taux d'erreur suffisamment bas pour estimer qu'aucune erreur ne se produise en utilisation normale. Toutefois, ces systèmes ne sont pas infaillibles. Avec une utilisation intensive, les gros utilisateurs tels que les flottes de serveurs font face à ces erreurs de manière courante [61]. Ces codes peuvent également être utilisés pour les OxRAM tels que dans [12]. Il est cependant utile de constater que ces codes de correction d'erreurs se reposent sur des hypothèses particulières concernant la statistique des erreurs. En particulier, une forte corrélation spatiale ou temporelle des erreurs peut drastiquement diminuer leur efficacité. Il peut alors être utile de développer des codes particuliers [62] se basant sur des hypothèses ou ayant des objectifs différents. Bien que les **codes correcteurs d'erreurs** émanent principalement du monde numérique, l'essor des paradigmes de calcul analogique a motivé le développement d'équivalents analogiques [63] [64] [65]. Il est également intéressant de

constater que la lecture différentielle de deux éléments mémoires OxRAM programmés dans un état opposé présente une diminution du taux d'erreur équivalente à celle d'un **code correcteur d'erreur** ayant le même niveau de redondance [39].

II.3 Méthode d'écriture

Dans le cas d'utilisation en tant que mémoire, une écriture précise des niveaux de conductance permet une diminution du taux d'erreur ou, alternativement, une augmentation de la densité d'état réalisable. De plus, comme nous l'avons vu dans le Chapitre I.3, de nombreuses autres applications de l'OxRAM profitent ou même dépendent d'une écriture précise du niveau de conductance de la cellule. L'écriture est donc un point central de la performance d'une OxRAM. Nous commençons par nous intéresser aux différentes méthodes d'écriture utilisées dans la littérature.

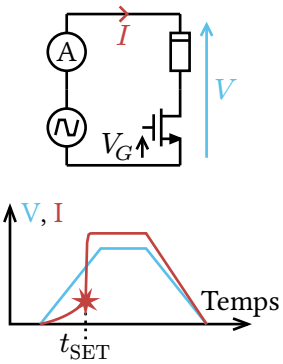
Une bonne méthode d'écriture cherche d'abord à améliorer la prédictibilité, l'endurance, la rétention, l'uniformité ainsi que la fenêtre de lecture (définies dans la partie Chapitre II.2.3). En plus de ces critères relatifs à la cellule, des critères propres à l'écriture peuvent être considérés.

1. **Durée de l'écriture**
   La durée de l'écriture est le temps nécessaire à l'écriture de la mémoire, du début, au moment où le circuit d'écriture est à nouveau disponible pour une autre écriture. Cette durée limite le débit d'écriture par circuit d'écriture. Un plus grand débit par circuit peut permettre de réduire leur nombre et ainsi augmenter la densité. La durée de l'écriture est également liée à la quantité de cache nécessaire dans un système mémoire complet, indépendamment du nombre de circuits d'écriture.
2. **Énergie d'écriture**
   L'énergie d'écriture est l'énergie consommée par le circuit d'écriture lors d'une unique écriture. Cette énergie peut limiter le nombre d'écritures effectuées simultanément, et/ou limiter la durée de vie d'un dispositif sur batterie. Dans le cas des applications de calculs nécessitant une écriture, cette énergie vient réduire la performance énergétique du système alors qu'elle est le principal critère que l'on cherche à optimiser dans ce cas.
3. **Sensibilité aux variations du circuit**
   La sensibilité aux variations du circuit correspond à la sensibilité de la méthode d'écriture à toute dérive des composants du circuit. Cela peut être la dérive d'une référence, d'une impédance de ligne, d'un autre circuit (par exemple le circuit de lecture lorsqu'il est également utilisé pour l'écriture). Cette sensibilité peut limiter l'intégration à grande échelle de la méthode, ou sur-contraindre les autres éléments du circuit.
4. **Sensibilité aux variations des cellules entre elles**
   La sensibilité aux variations des cellules entre elles est la sensibilité de la méthode d'écriture aux différences de caractéristiques d'une cellule à une autre. En particulier, elle est le pendant de l'uniformité et quantifie la capacité de la méthode d'écriture à tolérer et limiter le manque apparent d'uniformité des cellules.
5. **Densité**
   La densité est l'inverse de la surface requise pour l'implémentation du circuit d'écriture, ainsi que l'évolution de cette surface lorsque plusieurs circuits d'écriture sont intégrés (une partie du circuit peut ne pas avoir besoin d'être répétée). Elle participe à la densité globale du système mémoire.

Il est important de rappeler que la cellule 1T1R est constituée d'un élément mémoire OxRAM et d'un transistor MOSFET connecté en série. Lors de l'écriture, sauf mention d'une autre limite de courant, une tension adaptée est appliquée à la grille du transistor de la cellule qui est ainsi utilisée comme limiteur de courant. Nous considérons principalement le cas du SET, et discuterons des différences importantes avec celui du RESET et du FORMING à la fin de ce chapitre.

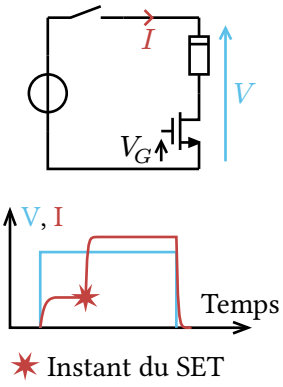
II.3.1 Quasi-statique

La méthode d'écriture quasi-statique consiste à soumettre la cellule à une tension croissante, puis décroissante, avec une pente relativement faible par rapport au temps caractéristique de la cellule. Elle est principalement utilisée pour mesurer une caractéristique quasi-statique de la cellule en observant le courant fourni à la cellule pendant ce processus. Le courant est limité pour ne pas détruire la cellule. Elle a l'intérêt d'être facilement réalisable en laboratoire à l'aide de source de tension externe sans avoir de problème d'impédance de ligne. Elle permet également d'obtenir une importante quantité d'informations avec la mesure de cette caractéristique. Elle n'est cependant pas utilisée dans les systèmes intégrés car elle est lente, et elle n'est pas la plus simple. Une source de tension est généralement utilisée, mais une source de courant peut également l'être [66] [67]. La source de courant est alors limitée en tension.



II.3.2 Impulsion à tension constante à durée déterminée

Cette méthode consiste à appliquer une impulsion de tension de durée déterminée sur la cellule. Cette méthode est très intégrable car elle ne nécessite qu'une source de tension constante et un interrupteur. La médiane de conductance Γ_{LRS} atteinte après l'écriture est proportionnelle au courant de limitation I_c^{eff}, et est faiblement dépendante de la durée de l'impulsion, si elle est assez longue, ou de l'amplitude de l'impulsion V_{SET} (comme décrit dans le Chapitre II.2.2). Cette méthode permet donc un contrôle simple du niveau médian de conductance atteint en jouant sur la limitation de courant.



L'amplitude de l'impulsion V_{SET} influence principalement le temps de SET, mais peut également être reliée au dépassement du courant (différence entre I_c et I_c^{eff}) provoqué par les capacités parasites lorsqu'elle est assez élevée. Elle influence également indirectement le courant de limitation I_c du fait de l'impédance finie du limiteur de courant.

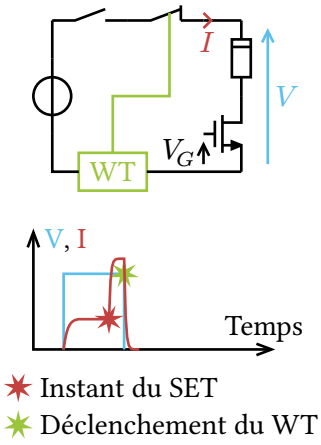
L'utilisation du transistor de la cellule 1T1R comme limiteur de courant, dit interne dans la suite, conduit à une forte sensibilité aux variations du circuit[22]. Certaines implémentations utilisent donc un limiteur de courant, dit externe, déporté en bordure de matrice. Cela permet d'augmenter la taille des transistors et donc de réduire leur variabilité de I_c, mais également d'utiliser le même limiteur pour un ensemble de cellules, ce qui réduit encore plus la variabilité perçue (puisque toutes les cellules utilisant le même limiteur verront le même I_c). Il est cependant nécessaire de prêter attention à la capacité parasite de ligne entre la cellule et le limiteur, qui réduit la précision dynamique de la limitation de courant au niveau de la cellule, cela se traduit par une différence entre I_c et I_c^{eff} accrue.

La consommation énergétique de cette méthode est principalement liée à la durée de l'impulsion, mais dépend également fortement du moment ou la transition entre HRS et LRS se produit réellement. En effet, avant la transition, le courant est faible car la résistance de la cellule est forte, mais après la transition, il est égal à la limitation de courant, qui est beaucoup plus élevée. Comme le moment de la transition est aléatoire, la durée doit être choisie assez grande pour assurer la transition de la cellule la plus lente à écrire. Ainsi une grande quantité d'énergie est dépensée après la transition alors qu'elle est, à première vue, inutile.

[22] Les transistors intégrés, même lorsqu'ils sont positionnés au plus proche, présentent des différences de caractéristique, notamment de courant de saturation. Cette différence est d'autant plus forte en moyenne que la surface occupée par les transistors est faible.

II.3.3 Impulsion à tension constante à durée adaptée

Pour résoudre ce problème de surconsommation d'énergie, il a été proposé de détecter la transition du niveau de conductance de la cellule pour mettre fin au plus vite à l'écriture [25]. Ainsi, la durée de l'impulsion est adaptée automatiquement à chaque cellule, mais également à chaque réalisation (dans le sens probabiliste) de l'écriture sur une cellule. Les méthodes d'écriture incorporant cette idée sont marquées dans la littérature par l'appellation « terminaison d'écriture » (« Write Terminaison »).



Cette méthode est plus complexe à mettre en place car elle nécessite un système de détection du changement d'état. La méthode habituelle est de détecter le moment où le limiteur de courant commence à limiter le courant. Elle se combine donc particulièrement bien avec les implémentations utilisant un limiteur de courant externe à la cellule.

Cet arrêt anticipé de l'impulsion permet de réduire considérablement l'énergie consommée, en moyenne, par l'écriture. Une augmentation de l'endurance est également attribuée à cette méthode [25] du fait d'une moindre sur-écriture des cellules qui favoriserait leur vieillissement rapide.

Certains retours d'expérience notent que l'arrêt au plus tôt de l'impulsion réduirait la rétention de la cellule [59]. Pour palier à ce défaut, il a été proposé de rajouter un délai volontaire entre la détection de la transition et l'arrêt de l'impulsion [11] [10].

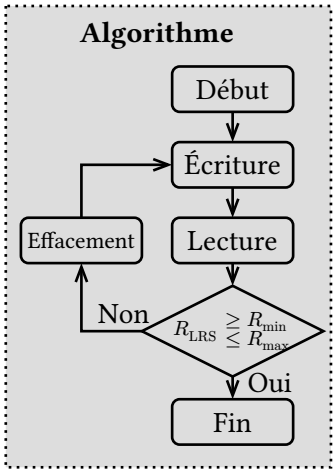
II.3.4 Écriture avec vérification

Les distributions de conductivités obtenues avec les méthodes précédentes peuvent être trop larges pour de nombreuses applications. Cela est particulièrement vrai lorsque l'on souhaite avoir une forte densité d'état, ou lorsque l'information doit être stockée sous forme analogique. En effet, il est alors nécessaire de réduire au maximum l'incertitude sur la conductance obtenue pour avoir le plus de niveaux distincts possibles, ou le moins de bruit possible dans le cas analogique.

Pour améliorer ces distributions, on peut se tourner vers des méthodes d'écriture incorporant une vérification. De façon générale, l'idée est d'effectuer une lecture après chaque opération d'écriture. L'information obtenue peut ensuite être utilisée pour agir de différente manière en fonction des cas, on parle alors d'algorithme de vérification. Ces méthodes créent un compromis entre durée d'écriture et prédictibilité.

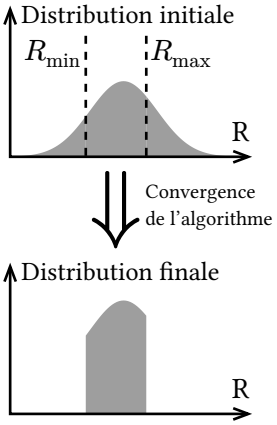
Avant de détailler les différentes approches, précisons que toutes les méthodes de lecture n'offrent pas la même quantité d'informations. Dans le meilleur des cas, une image précise de la conductance de la cellule peut être obtenue, mais dans d'autres cas, seule la position de la conductance par rapport à un seuil est obtenue. L'information disponible contraint les algorithmes de vérification qu'il est possible de mettre en œuvre. Généralement, plus une méthode de lecture fournit d'informations, plus elle requerra du temps et de la surface. Il faut donc prêter attention à la quantité d'informations requise par un algorithme de vérification lors de son développement.

a) Ré-écriture avec vérification



Une première méthode de vérification consiste, de façon imagée, à retenter sa chance à chaque fois que l'écriture échoue ou ne convient pas au critère fixé. Pour ce faire, le cellule est d'abord effacée, puis ré-écrite.

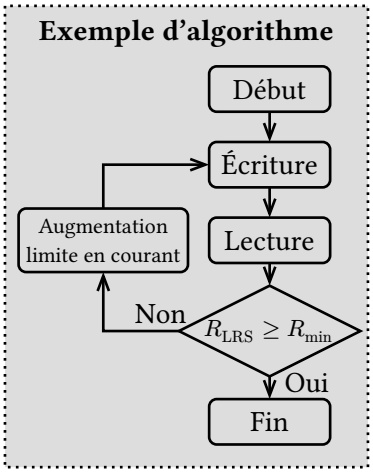
Cette méthode vient donc déformer la distribution de conductance obtenue avec la méthode d'écriture en rejetant tous les points ne respectant pas le critère donné. Dans la limite ou chaque tentative d'écriture est indépendante et identiquement distribuée aux précédentes, le nombre d'écritures à effectuer pour satisfaire le critère suit une loi géométrique dont le paramètre est égal à la probabilité de satisfaire le critère à la première tentative. L'espérance est alors l'inverse de cette probabilité. Cette méthode requiert donc d'autant plus d'essais que l'on souhaite réduire la distribution de conductance de sortie. Il y a donc un compromis entre amélioration de la prédictibilité et le nombre de tentatives, et donc la durée d'écriture, ainsi que l'énergie d'écriture et également l'endurance effective.



Cette méthode n'améliore pas la sensibilité aux variations du circuit ou des cellules entre elles, elle y est même très sensible. Si la méthode d'écriture utilisée présente une distribution de conductance résultante très variable selon la cellule visée, ou que l'écriture échoue systématiquement sur certaines cellules qui seraient « bloquées » dans un état, le nombre de tentatives peut augmenter rapidement au-dessus de l'estimation précédente. Un nombre de tentatives maximum est donc utilisé pour prévenir ces cas.

Cette méthode a un coût d'implémentation faible. En effet, au-delà du circuit d'écriture qui devrait déjà être présent, seule une lecture à deux seuils et un peu de logique sont requises.

b) Impulsions adaptatives avec vérification



La précédente méthode avait comme inconvénient d'être sensible aux variations du circuit et des cellules entre elles. Pour résoudre cela, une autre méthode de vérification est généralement utilisée.

L'objectif est d'avoir une écriture adaptative pour maximiser nos chances d'écrire chaque cellule à la conductance que l'on souhaite. L'idée est toujours d'effectuer plusieurs écritures et lectures successives. Mais cette fois, les écritures s'enchaînent sans effacement entre elles et chaque écriture voit ses paramètres adaptés selon l'algorithme de vérification. Généralement, l'algorithme consiste à commencer par des impulsions d'écriture de faible amplitude, ou avec une limite en courant faible, puis d'augmenter ces paramètres petit à petit. Éventuellement, la polarité de l'impulsion peut également être alternée en cas de dépassement. Cette méthode est très utilisée pour les applications souhaitant une forte densité d'état, souvent avec la perspective applicative des réseaux de neurones [68] [69] [27] [37] [36] [32] [70] [71]. Mais elle est également utilisée pour réduire le taux d'erreur de module mémoire [12].

Cette méthode permet d'améliorer grandement l'uniformité, et comme la précédente, la prédictibilité. Il y a généralement un compromis entre cette amélioration et le nombre de tentatives, et donc

la durée d'écriture, ainsi que l'énergie d'écriture. Plus les pas d'augmentation des paramètres sont grands, moins le nombre de tentatives sera élevé, mais moins l'adaptabilité sera grande. Cette méthode permet également d'absorber les variations dues aux processus de fabrication sans avoir recours à de la calibration [9].

Cette méthode est plus contraignante à mettre en place. D'un point de vue de la lecture, la contrainte dépend de la quantité d'informations que l'on souhaite avoir pour l'algorithme. Pour la plupart seulement deux seuils sont requis, comme pour la méthode précédente. L'augmentation des contraintes se trouve principalement sur le module d'écriture. En effet, l'amplitude des impulsions, de la limite en courant, ou des deux doit pouvoir être adaptée avec la granularité demandée par l'algorithme.

On peut noter que l'on peut difficilement accélérer cette écriture en gardant en mémoire les paramètres finaux car ils changent à chaque cycle, mais également car cela représenterait une quantité d'informations plus importante que celle stockable dans les éléments mémoires OxRAM. Cela serait donc contre-productif. Par contre, des paramètres plus globaux peuvent être enregistrés comme le fait [12], pour trouver le meilleur paramétrage de l'algorithme en fonction de la puce.

c) Limitations

Avec ces méthodes, on pourrait imaginer pouvoir atteindre n'importe quelle distribution avec assez d'étapes. Mais le phénomène de relaxation vient contrarier ce plan. Ce phénomène à tendance à régénérer les distributions obtenues avec une écriture sans vérification [72] [69]. L'amplitude des variations est plus importante peu de temps après l'écriture, et s'atténue avec le temps. Ajouter un délai entre l'écriture et la lecture de vérification permet d'améliorer les distributions obtenues [27]. Cependant, cela aggrave le compromis que l'on avait déjà identifié avec la durée d'écriture.

La cause de la relaxation sur le temps court est attribuée soit au déplacement d'ions oxygène générés en excès lors de l'écriture, en particulier par le dépassement de courant dû aux capacités parasites [70], soit par la relaxation des trous d'oxygène constituant le filament [58] [72]. Il a été proposé d'utiliser des impulsions triangulaires, ou des rampes de tension pour diminuer la sur-génération d'ions oxygène [70]. Il a également été observé que la relaxation était moins forte après une impulsion polarisée dans le sens du RESET que dans celui du SET. Des optimisations aux niveaux du matériau peuvent également être envisagées pour stabiliser les trous [58].

Au-delà de la relaxation, les bruits et autres variations temporels de la conductance discutés à la fin de la partie Chapitre II.2.2 peuvent perturber la vérification et limiter le support minimal des distributions atteignables.

II.3.5 Forming

La première écriture de la vie de l'élément mémoire OxRAM est un peu particulière. Comme il n'y a pas encore de filament partiel dans l'oxyde, il est nécessaire d'appliquer une différence de potentiel plus importante que pendant les SET suivants pour générer un champ électrique suffisamment fort dans l'oxyde. La problématique est donc la même, mais avec des différences de potentiels plus élevées.

Comme la tension est plus élevée, le dépassement de courant dû aux capacités parasites est plus important. On préfère donc limiter le courant pendant la FORMING à une valeur plus faible que celle ciblée pour les SET à venir. Néanmoins, il a été rapporté que la condition de FORMING n'a que peu d'influence sur les caractéristiques de l'élément mémoire OxRAM après un nombre de cycles important [73] [74].

La tension nécessaire au FORMING est un frein à la diminution de la dimension des éléments mémoires OxRAM. En effet, la tension nécessaire est inversement proportionnelle à la surface de l'élément mémoire OxRAM [75]. Cette tension pose problème lors de l'utilisation de nœuds CMOS avancés pour constituer les circuits périphériques car leurs transistors les plus fins ne supportent pas ces tensions.

Pour diminuer cette tension de FORMING, dans le cas des OxRAMs à base d'oxyde d'hafnium, il a été proposé d'effectuer une étape d'implantation de silicium [76] [77] dans l'oxyde. De plus, des études sur les transistors de certaines technologies ont montré que leur utilisation à plus haute tension que prévu sur des temps courts ne crée pas de dégradation de leur caractéristique [60]. Comme le FORMING n'est réalisé qu'une seule fois, les tensions élevées qui lui sont associées ne sont pas nécessaires sur un temps long. Il y a donc une marge de manœuvre sur cet aspect. Toutefois, cela demande des études de fiabilité des transistors et des circuits périphériques utilisés hors de leur zone de fonctionnement garanties. Les modèles fournis par le fabricant ne sont alors pas suffisants, et des expérimentations spécifiques doivent être réalisées.

II.3.6 Type de source utilisée pour l'écriture

a) Écriture avec une source de tension

Comme le champ électrique est l'élément moteur de la réaction d'écriture [49], il est naturel de souhaiter imposer une tension sur la cellule pour l'écrire. Imposer une tension, comme le fait une source de tension, est également le moyen le plus habituel d'imposer une excitation électrique. L'utilisation d'une source de tension est donc le moyen le plus couramment utilisé pour écrire les OxRAMs. C'est d'ailleurs celui considéré de facto dans la précédente partie.

Comme décrit précédemment, ce mode d'écriture requiert une limitation de courant pour le SET, sans quoi l'oxyde serait trop dégradé pour que la réaction soit réversible. Il a été identifié que le courant de limitation I_c était le principal paramètre définissant le niveau de conductance Γ_{LRS} atteint [50] [52]. Le courant circulant dans l'OxRAM semble donc être une grandeur importante à contrôler. Plutôt que de fixer la tension, nous pourrions essayer de fixer le courant avec une source de courant.

b) Écriture avec une source de courant

Une source de courant fixe le courant qui la traverse. Cela a pour effet de moduler la tension à ses bornes. La tension est alors une conséquence, et sa valeur est définie par le reste du circuit. La source de courant est l'exact dual de la source de tension, qui elle fixe la tension et laisse le courant être défini par le reste du circuit.

Lors de l'utilisation d'une source de tension, le SET est soudain, c'est à dire que la durée de la transition est courte (moins de 100 ps), assez pour être difficilement observable. Les deux grandeurs principales régulant la vitesse de la réaction sont la température et le champ électrique local dans l'oxyde. Pendant le SET, la résistance de la cellule ainsi que l'épaisseur d'oxyde effective diminuent. Ainsi, à tension fixée, la puissance dissipée dans la cellule, et donc la température, augmente en même temps que la réaction progresse. De même, comme le filament croit, la distance effective d'oxyde diminue, et le champ électrique local augmente en conséquence. Il y a donc une rétroaction positive. Plus la réaction progresse, plus elle est favorisée. Cela explique le caractère soudain du SET, que l'on ne semble pas pouvoir ralentir, et cela réaffirme la nécessité d'une limitation en courant pour mettre fin à cette accélération.

A l'inverse, en RESET, la transition observée est progressive. En conduisant la même analyse, on constate que la rétroaction est cette fois négative. En effet, la résistance augmente, ce qui diminue la puissance et donc la température. De plus le filament se dissout, augmentant l'épaisseur effective d'oxyde et diminuant le champ électrique local. Cette rétro-action négative explique l'auto-modération du RESET, où la tension d'excitation doit être continuellement augmentée pour continuer à faire progresser la réaction. Cela explique également la relation positive observée entre R_{HRS} et V_{RESET}. Plus V_{RESET} est élevée, plus l'épaisseur effective d'oxyde finale sera grande, et donc plus la résistivité le sera également. Cela augmentera également la tension nécessaire au prochain SET $V_{\mathrm{SET}}^{\mathrm{eff}}$.

L'utilisation d'une source de courant vient inverser ces considérations. En effet, à courant constant, la diminution de la résistance lors d'un SET conduit, au contraire, à une diminution de la puissance dissipée et à une augmentation lors d'un RESET. On pourrait donc s'attendre à un SET progressif et un RESET soudain. L'analyse sur le sens de variation du champ électrique local est cependant moins évidente dans cette situation. En effet, lorsque la résistance diminue, la tension diminue, néanmoins l'épaisseur effective d'oxyde diminue également. Ainsi la variation du champ électrique dépend du lien entre la variation de tension et la variation de l'épaisseur effective d'oxyde. Dans la littérature, des conclusions divergentes apparaissent. D'un côté, [66] observe un SET progressif et un RESET soudain lors de l'utilisation d'une source de courant. D'un autre côté, [67] observe un SET et un RESET tous deux soudains. La raison de cette disparité provient peut-être des échelles de temps mises en jeu[23]. Cependant, aucune information n'est donnée concernant la fréquence des points de mesure présentés dans ces deux articles.

Dans la littérature, on trouve quelques évocations de l'utilisation d'une source de courant pour programmer les OxRAM. Il y a les deux expériences d'écriture quasi-statique précédemment citées [66] [67] utilisant de l'équipement de laboratoire et des dispositifs isolés des autres. Elles ne sont donc pas forcement représentatives d'un circuit d'écriture intégré ainsi que d'une utilisation avec une matrice de cellules. Une implémentation intégrée de module mémoire utilisant une source de courant pour l'opération de SET est présentée dans [9]. Il y est dit que cela « permet un contrôle précis du filament », mais cet aspect n'est pas développé au-delà.

L'écriture utilisant une source de courant semble donc encore bien mystérieuse. C'est pourquoi les deux prochains chapitres lui seront destinés. Dans le Chapitre III, nous essayerons cette méthode d'écriture sur un nombre limité de dispositifs 1R isolés entre eux et écrits par une source de courant co-intégrée. Nous y observerons l'influence importante de la capacité parasite qui tend à limiter notre capacité à imposer le courant. Toutefois, l'architecture du circuit, qui n'avait pas été conçue pour cette expérience, ne permettra pas d'obtenir une statistique satisfaisante, ni d'effectuer une étude comparative entre écriture en courant et écriture en tension. Les travaux présentés dans le Chapitre IV tenteront de résoudre ces problèmes. Nous discuterons d'un circuit intégré spécifiquement développé pour l'étude de l'écriture en courant. Il permettra l'obtention de statistiques sur 16 mille (2^{14}) cellules 1T1R par puce, ainsi que la possibilité de comparer les écritures en courant et en tension directement sur la même puce.

c) Écriture avec une source capacitive

Dans la partie Chapitre II.3.3, nous avons discuté d'une technique d'écriture en SET ayant pour objectif de mettre fin à l'excitation dès que la transition était détectée. L'intérêt est de diminuer l'énergie inutile consommée après la transition. Cependant, l'application de cette technique requiert l'ajout d'un circuit de détection. Il serait préférable que cet arrêt de l'excitation se produise naturellement. On peut obtenir ce comportement en utilisant une source capacitive.

Une telle source consiste en une capacité pré-chargée à une tension. Lorsque cette source est connectée à la cellule, un courant circule et décharge la capacité. Lors d'un SET, la résistance initiale de la cellule est élevée. La décharge est donc relativement lente. Dès que la transition se produit, la résistance de la cellule chute et la décharge s'accélère en conséquence. Au bout d'un temps fini, l'intégralité de la capacité se sera déchargée. Comme la différence de résistance entre les deux états de la cellule est grande (au moins un ordre de grandeur), la différence en vitesse de décharge l'est également. En utilisant une taille de capacité adaptée, on pourrait espérer avoir une décharge initiale suffisamment

[23]La notion de transition « soudaine » est définie relativement au temps entre deux observations successives. En effet, une transition est considérée soudaine lorsqu'aucune observation n'est obtenue pendant celle-ci. Au contraire, elle n'est pas « soudaine » lorsque plusieurs observations sont obtenues pendant celle-ci. La fréquence d'observation est donc un élément clé dans cette définition.

lente pour parvenir à écrire la cellule, et une décharge finale assez rapide pour émuler un arrêt de l'excitation peu après la « détection » de la transition.

Cette manière d'écrire a déjà été expérimentée dans la littérature [78] [79]. L'intérêt énergétique face à l'utilisation d'une source de tension y est développé. La taille minimale de la capacité nécessaire est également recherchée dans [79], et est expérimentalement établie à l'ordre de quelques centaines de picofarad. Cette quantité est importante et à la limite de l'intégrable. Cependant, dans les expériences de [78] et [79], les cellules ne sont pas co-intégrées avec la source (la capacité pré-chargée), ainsi une forte capacité parasite se trouve entre les cellules écrites et la source. Cela conduit nécessairement à un partage de charge important ayant pour effet de réduire la tension effective de pré-charge et donc de réduire l'énergie disponible pour l'écriture. On peut donc s'attendre à ce qu'une solution intégrée nécessite une capacité plus faible.

L'exploration de cette idée d'utiliser une source capacitive pour l'écriture des OxRAM est le sujet du Chapitre V. Il traite d'une manière plus détaillée l'aspect théorique, puis il présente le développement d'un circuit intégré conçu spécifiquement pour l'expérimentation de cette écriture.

Bibliographie

[9] P. Jain *et al.*, « 13.2 A 3,6 Mb 10,1 Mb/mm2 Embedded Non-Volatile ReRAM Macro in 22 nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0,5 V with Sensing Time of 5 ns at 0,7 V », in *2 019 IEEE International Solid- State Circuits Conference - (ISSCC)*, févr. 2019, p. 212-214. doi: 10.1109/ISSCC.2019.8662393.

[10] J. Yang *et al.*, « 24.2 A 14 nm-FinFET 1 Mb Embedded 1T1R RRAM with a 0.022µm2 Cell Size Using Self-Adaptive Delayed Termination and Multi-Cell Reference », in *2 021 IEEE International Solid- State Circuits Conference (ISSCC)*, févr. 2021, p. 336-338. doi: 10.1109/ISSCC42613.2021.9365945.

[11] C.-C. Chou *et al.*, « An N40 256K×44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance », in *2 018 IEEE International Solid - State Circuits Conference - (ISSCC)*, févr. 2018, p. 478-480. doi: 10.1109/ISSCC.2018.8310392.

[12] B. Giraud *et al.*, « Benefits of Design Assist Techniques on Performances and Reliability of a RRAM Macro », in *2 023 IEEE International Memory Workshop (IMW)*, mai 2023, p. 1-4. doi: 10.1109/IMW56887.2023.10145984.

[25] W.-H. Chen *et al.*, « A 16 Mb dual-mode ReRAM macro with sub-14 ns computing-in-memory and memory functions enabled by self-write termination scheme », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 28. doi: 10.1109/IEDM.2017.8268468.

[27] E. Esmanhotto *et al.*, « Experimental demonstration of Single-Level and Multi-Level-Cell RRAM-based In-Memory Computing with up to 16 parallel operations », in *2 022 IEEE International Reliability Physics Symposium (IRPS)*, mars 2022, p. P8–1–P8–4. doi: 10.1109/IRPS48227.2022.9764474.

[32] A. Glukhov *et al.*, « Statistical model of program/verify algorithms in resistive-switching memories for in-memory neural network accelerators », in *2 022 IEEE International Reliability Physics Symposium (IRPS)*, mars 2022, p. 3. doi: 10.1109/IRPS48227.2022.9764497.

[36] V. Milo *et al.*, « Accurate Program/Verify Schemes of Resistive Switching Memory (RRAM) for In-Memory Neural Network Circuits », *IEEE Transactions on Electron Devices*, vol. 68, nº 8, p. 3832-3837, août 2021, doi: 10.1109/TED.2021.3089995.

[37] S. Yu, « Neuro-inspired computing with emerging nonvolatile memorys », *Proceedings of the IEEE*, vol. 106, nº 2, p. 260-285, févr. 2018, doi: 10.1109/JPROC.2018.2790840.

[39] M. Bocquet *et al.*, « In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks », in *2 018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 20. doi: 10.1109/IEDM.2018.8614639.

[45] K.-J. Zhou *et al.*, « Abnormal High Resistive State Current Mechanism Transformation in Ti/HfO2/TiN Resistive Random Access Memory », *IEEE Electron Device Letters*, vol. 41, nº 2, p. 224-227, févr. 2020, doi: 10.1109/LED.2019.2961408.

[46] R. Waser, R. Dittmann, G. Staikov, et K. Szot, « Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges », *Advanced Materials*, vol. 21, nº 25–26, p. 2632-2663, 2009, doi: 10.1002/adma.200900375.

[47] H.-S. P. Wong *et al.*, « Metal–Oxide RRAM », *Proceedings of the IEEE*, vol. 100, nº 6, p. 1951-1970, juin 2012, doi: 10.1109/JPROC.2012.2190369.

[48] F. M. Puglisi, N. Zagni, L. Larcher, et P. Pavan, « Random Telegraph Noise in Resistive Random Access Memories: Compact Modeling and Advanced Circuit Design », *IEEE Transactions on Electron Devices*, vol. 65, nº 7, p. 2964-2972, juill. 2018, doi: 10.1109/TED.2018.2833208.

[49] D. Ielmini, F. Nardi, et S. Balatti, « Evidence for Voltage-Driven Set/Reset Processes in Bipolar Switching RRAM », *IEEE Transactions on Electron Devices*, vol. 59, nº 8, p. 2049-2056, août 2012, doi: 10.1109/TED.2012.2199497.

[50] F. Nardi, S. Larentis, S. Balatti, D. C. Gilmer, et D. Ielmini, « Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part I: Experimental Study », *IEEE Transactions on Electron Devices*, vol. 59, nº 9, p. 2461-2467, sept. 2012, doi: 10.1109/TED.2012.2202319.

[51] Z.-Q. Wang *et al.*, « Cycling-induced degradation of metal-oxide resistive switching memory (RRAM) », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 7. doi: 10.1109/IEDM.2015.7409649.

[52] D. Ielmini, « Resistive switching memories based on metal oxides: mechanisms, reliability and scaling », *Semiconductor Science and Technology*, vol. 31, nº 6, p. 63002, juin 2016, doi: 10.1088/0268-1242/31/6/063002.

[53] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, et D. Ielmini, « Statistical Fluctuations in HfOx Resistive-Switching Memory: Part I - Set/Reset Variability », *IEEE Transactions on Electron Devices*, vol. 61, nº 8, p. 2912-2919, août 2014, doi: 10.1109/TED.2014.2330200.

[54] A. Grossi *et al.*, « Fundamental variability limits of filament-based RRAM », in *2 016 IEEE International Electron Devices Meeting (IEDM)*, déc. 2016, p. 4. doi: 10.1109/IEDM.2016.7838348.

[55] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, et D. Ielmini, « Statistical Fluctuations in HfOx Resistive-Switching Memory: Part II—Random Telegraph Noise », *IEEE Transactions on Electron Devices*, vol. 61, nº 8, p. 2920-2927, août 2014, doi: 10.1109/TED.2014.2330202.

[56] S. Ambrogio, S. Balatti, V. McCaffrey, D. C. Wang, et D. Ielmini, « Noise-Induced Resistance Broadening in Resistive Switching Memory—Part I: Intrinsic Cell Behavior », *IEEE Transactions on Electron Devices*, vol. 62, nº 11, p. 3805-3811, nov. 2015, doi: 10.1109/TED.2015.2475598.

[57] S. Ambrogio, S. Balatti, V. McCaffrey, D. C. Wang, et D. Ielmini, « Noise-Induced Resistance Broadening in Resistive Switching Memory—Part II: Array Statistics », *IEEE Transactions on Electron Devices*, vol. 62, nº 11, p. 3812-3819, nov. 2015, doi: 10.1109/TED.2015.2477135.

[58] L. Reganaz *et al.*, « Elucidating Postprogramming Relaxation in Multilevel Cell-Resistive Random Access Memory by Means of Experimental and Kinetic Monte Carlo Simulation Data », *physica status solidi (a)*, nº n/a, p. 2100753, doi: 10.1002/pssa.202100753.

[59] C. Y. Chen *et al.*, « Programming-conditions solutions towards suppression of retention tails of scaled oxide-based RRAM », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 10. doi: 10.1109/IEDM.2015.7409671.

[60] J. Sandrini *et al.*, « OxRAM for embedded solutions on advanced node: scaling perspectives considering statistical reliability and design constraints », in *2 019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 30. doi: 10.1109/IEDM19573.2019.8993484.

[61] J. Meza, Q. Wu, S. Kumar, et O. Mutlu, « Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field », in *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, juin 2015, p. 415-426. doi: 10.1109/DSN.2015.57.

[62] H. Tang et J. Park, « Unequal-Error-Protection Error Correction Codes for the Embedded Memories in Digital Signal Processors », *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, nº 6, p. 2397-2401, juin 2016, doi: 10.1109/TVLSI.2015.2497368.

[63] C. Li, R. M. Roth, C. Graves, X. Sheng, et J. P. Strachan, « Analog error correcting codes for defect tolerant matrix multiplication in crossbars », in *2 020 IEEE International Electron Devices Meeting (IEDM)*, déc. 2020, p. 36. doi: 10.1109/IEDM13553.2020.9371978.

[64] R. M. Roth, « Analog Error-Correcting Codes », *IEEE Transactions on Information Theory*, vol. 66, nº 7, p. 4075-4088, juill. 2020, doi: 10.1109/TIT.2020.2977918.

[65] M. Jiang, R. Mao, J. P. Strachan, et C. Li, « Defect tolerant in-memory analog computing with CMOS-integrated nanoscale crossbars: Invited », in *2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, nov. 2021, p. 1-4. doi: 10.1109/ICECS53924.2021.9665649.

[66] B. Chen *et al.*, « A Novel Operation Scheme for Oxide-Based Resistive-Switching Memory Devices to Achieve Controlled Switching Behaviors », *IEEE Electron Device Letters*, vol. 32, nº 3, p. 282-284, mars 2011, doi: 10.1109/LED.2010.2101577.

[67] W. Lian *et al.*, « Improved Resistive Switching Uniformity in $ \hboxCu/HfO_2/\hboxPt$ Devices by Using Current Sweeping Mode », *IEEE Electron Device Letters*, vol. 32, nº 8, p. 1053-1055, août 2011, doi: 10.1109/LED.2011.2157990.

[68] Y. Luo, X. Han, Z. Ye, H. Barnaby, J.-S. Seo, et S. Yu, « Array-Level Programming of 3-Bit per Cell Resistive Memory and Its Application for Deep Neural Network Inference », *IEEE Transactions on Electron Devices*, vol. 67, nº 11, p. 4621-4625, nov. 2020, doi: 10.1109/TED.2020.3015940.

[69] E. Esmanhotto *et al.*, « High-Density 3D Monolithically Integrated Multiple 1T1R Multi-Level-Cell for Neural Networks », in *2 020 IEEE International Electron Devices Meeting (IEDM)*, déc. 2020, p. 36. doi: 10.1109/IEDM13553.2020.9372019.

[70] Y. Feng *et al.*, « Improvement of State Stability in Multi-Level Resistive Random-Access Memory (RRAM) Array for Neuromorphic Computing », *IEEE Electron Device Letters*, vol. 42, nº 8, p. 1168-1171, août 2021, doi: 10.1109/LED.2021.3091995.

[71] B. Q. Le *et al.*, « RADAR: A Fast and Energy-Efficient Programming Technique for Multiple Bits-Per-Cell RRAM Arrays », *IEEE Transactions on Electron Devices*, vol. 68, nº 9, p. 4397-4403, sept. 2021, doi: 10.1109/TED.2021.3097975.

[72] A. Fantini *et al.*, « Intrinsic program instability in HfO2 RRAM and consequences on program algorithms », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 7. doi: 10.1109/IEDM.2015.7409648.

[73] C.-Y. Wu *et al.*, « Emerging Memory RRAM Embedded in 12FFC FinFET Technology for industrial Applications », in *2 023 International Electron Devices Meeting (IEDM)*, déc. 2023, p. 1-4. doi: 10.1109/IEDM45741.2023.10413775.

[74] C. Peters, F. Adler, K. Hofmann, et J. Otterstedt, « Reliability of 28 nm embedded RRAM for consumer and industrial products », in *2 022 IEEE International Memory Workshop (IMW)*, mai 2022, p. 1-3. doi: 10.1109/IMW52921.2022.9779300.

[75] S. Amer, M. S. Hasan, et G. S. Rose, « Analysis and Modeling of Electroforming in Transition Metal Oxide-Based Memristors and Its Impact on Crossbar Array Density », *IEEE Electron Device Letters*, vol. 39, nº 1, p. 19-22, janv. 2018, doi: 10.1109/LED.2017.2778639.

[76] M. Barlas *et al.*, « Improvement of HfO2 based RRAM array performances by local Si implantation », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 14. doi: 10.1109/IEDM.2017.8268392.

[77] L. Grenouillet *et al.*, « 16kbit 1T1R OxRAM arrays embedded in 28 nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors », in *2 021 IEEE International Memory Workshop (IMW)*, mai 2021, p. 1-4. doi: 10.1109/IMW51353.2021.9439607.

[78] S. Zhang *et al.*, « Programming resistive switching memory by a charged capacitor », *Applied Physics A*, vol. 102, nº 4, p. 1003-1007, mars 2011, doi: 10.1007/s00339-011-6320-5.

[79] P. Trotti, G. Pillonet, G. Molas, S. Oukassi, et E. Nowak, « Experimental Set-Up For Novel Energy Efficient Charge-based Resistive RAM (RRAM) Switching », in *2 020 IEEE 33rd International Conference on Microelectronic Test Structures (ICMTS)*, mai 2020, p. 1-5. doi: 10.1109/ICMTS48187.2020.9107936.

Chapitre III : Écriture en courant sur des dispositifs unitaires

III.1 Introduction

Comme nous l'avons vu dans le chapitre d'introduction précédent, la grandeur expérimentale ayant un effet le plus direct sur R_{LRS} est le courant maximal ayant traversé la cellule I_c^{eff}. Ainsi l'utilisation d'une source de courant permettant de fixer cette variable semble être une approche intéressante pour améliorer l'écriture des éléments mémoires OxRAM.

On trouve dans la littérature deux expériences utilisant des instruments de laboratoire explorant cette idée [66] [67] en réalisant une mesure de caractéristique quasi-statique avec une source de courant. Ces deux expériences ont l'inconvénient d'utiliser l'écriture quasi-statique qui n'est pas représentative des écritures utilisées par les implémentations intégrées. De plus, ces deux expériences font deux observations contraires l'une de l'autre. L'une observe une transition soudaine du niveau de résistance de R_{HRS} à R_{LRS}, comme c'est le cas avec une source de tension, alors que l'autre observe une transition lente et progressive, comme l'est le RESET avec une source de tension. La littérature contient également une présentation d'un circuit [9] faisant état de l'utilisation d'une méthode de SET en courant. Cependant, outre l'affirmation que cela « permet un contrôle précis du filament », cet aspect n'est pas discuté dans la publication.

Ainsi, l'utilisation d'une source de courant semble encore bien mystérieuse. C'est pourquoi nous souhaitons effectuer notre propre expérimentation. La littérature contient de nombreux modèles comportementaux de l'OxRAM [80] [48] [81] [32]. Néanmoins, ces modèles, bien que souvent fondés sur des principes physiques fondamentaux, requièrent une étape de calibration. Cette calibration apporte son lot de biais. Dans le cas de l'OxRAM, la calibration requiert l'écriture des éléments mémoires OxRAM. Comme la méthode standard d'écriture utilise une source de tension, les modèles sont basés sur cette hypothèse. Ainsi, il n'est pas évident qu'un modèle, même calibré, soit représentatif de la réalité lors de l'utilisation d'une source de courant. Nous nous tournons donc vers l'expérimentation.

Des méthodes d'écriture utilisant des sources de courant ont déjà été expérimentées en régime quasi-statique par [66] [67]. Nous souhaitons aller plus loin en expérimentant une écriture utilisant une source de courant co-intégrée aux éléments mémoires OxRAM pour explorer une situation plus proche du cas d'application réel. Dans ce chapitre, nous utiliserons un circuit intégré développé dans le cadre de la thèse de Paola Trotti. Ce circuit avait été développé pour l'expérimentation d'une écriture à durée limitée (« Write Terminaison »). N'ayant pas été développé spécifiquement pour cette expérience, ce circuit nous imposera certaines limitations que nous dépasserons dans le Chapitre IV avec un circuit spécifiquement développé pour nos expériences. En particulier, le circuit utilisé ici contient des

éléments mémoires OxRAM sans transistor de sélection. En conséquence, dans ce chapitre, nous ne considérons pas une cellule 1T1R, mais une cellule 1R (c'est-à-dire un élément mémoire OxRAM seul).

Les résultats présentés dans ce chapitre ont fait l'objet d'une publication dans la révue scientifique IEEE Electron Device Letters[24].

III.2 Écriture en courant

III.2.1 La relation entre R_{LRS} et I_c^{eff}

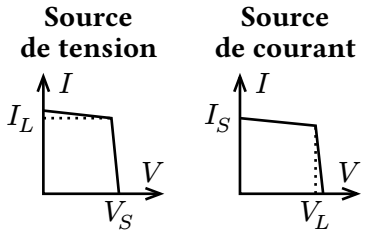
Le paramètre ayant l'influence la plus directe sur R_{LRS}, la résistance à l'état écrit, est I_c^{eff}, le courant maximal ayant traversé l'élément mémoire OxRAM pendant son écriture. Ainsi, l'amélioration de la prédictibilité de l'écriture en SET semble devoir passer par un meilleur contrôle de I_c^{eff}.

Dans la méthode standard utilisant une source de tension ainsi qu'une limitation en courant, ce contrôle est effectué au travers de I_c, le courant de limitation. Plusieurs effets peuvent venir réduire la précision. Premièrement, une différence entre I_c et I_c^{eff} peut apparaître à cause de courant de dépassement correspondant aux charges/décharges capacitives dues aux capacités des interconnexions présentes entre le limiteur de courant et l'élément mémoire OxRAM [51]. Ce phénomène est d'autant plus important que la dynamique de transition entre R_{HRS} et R_{LRS} est grande, et donc que V_{SET} est grand. Il est également d'autant plus important que la capacité d'interconnexion mise en jeu est grande. On a donc tout intérêt à rapprocher le circuit de limitation de courant de la cellule pour limiter la capacité. L'idéal d'après ce point de vue est d'utiliser la transistor de la cellule 1T1R comme limiteur. La capacité d'interconnexion effective correspond alors à la capacité du nœud connectant le transistor à l'élément mémoire OxRAM au sein de la cellule. Elle est donc par construction très faible.

Toutefois, le limiteur de courant utilisé a également une précision propre et limitée, qui peut dominer la précision globale de la cellule [54]. Dans le cas d'un circuit CMOS, cette précision est au premier ordre proportionnelle à la surface. Sous cet angle de vue, l'utilisation du transistor de la cellule 1T1R en tant que limiteur est une mauvaise idée, puisqu'il est nécessairement de petite taille pour des questions de densité. Ainsi, les implémentations tendent à utiliser des limiteurs de courant externe à la matrice. Dans ce cas, le problème de dépassement du courant de limitation peut être un problème car la capacité d'interconnexion est beaucoup plus forte.

Devant ce compromis, une solution semble d'explorer d'autres méthodes d'écriture. Avec le lien important existant entre R_{LRS} et I_c^{eff}, l'utilisation d'une source de courant semble être un bon candidat. A première vue, on pourrait espérer avoir un meilleur contrôle de I_c^{eff} et également ne plus faire face au même compromis.

III.2.2 Apparente équivalence entre sources avec limitation



En pratique, une source de courant est limitée en tension. En régime statique, que les sources et les limites soient parfaites ou qu'elles présentent une impédance, il y a équivalence entre source de tension limitée en courant et source de courant limitée en tension, comme illustré ci-contre par leur caractéristique respective. On pourrait donc *a priori* se dire que « l'écriture en courant » n'est rien d'autre qu'une écriture en tension déguisée. Cependant, en considérant le régime dynamique, on constate qu'il existe de réelles différences.

[24] Thomas Bauvent, Paola Trotti, Olivier Billoint, Jean-François Nodin, Yasser Moursy, Gabriel Molas, et Gaël Pillonnet. « RRAM Device Performances Under Capacitive-Enhanced Current Programming Scheme ». IEEE Electron Device Letters 44, nº 7 (juillet 2023): 1112-15. DOI: 10.1109/LED.2023.3274219.

III.2.3 Différences dues aux effets dynamiques

L'effet dynamique principal provient de la capacité d'interconnexion de la ligne reliant la source et la cellule. Si l'on prend en compte cette capacité, notée C_p dans la suite, on observe des différences importantes entre les deux types de source. La Figure 3.1 illustre ces différences.

Dans le cas de la source de tension, la source fixe la différence de potentiel aux bornes de C_p. Ainsi, pendant l'écriture, le courant dans C_p est nul, et les choses se passent comme si C_p n'était pas présente[25]. Dans le cas de la source de courant, la tension n'est pas fixée. Ainsi, le courant dans C_p dépend de la variation de la tension. A la connexion de la source de courant, la tension étant supposée initialement nulle, l'intégralité du courant passe au travers de la capacité. Elle se charge et la tension augmente. Le courant commence alors à circuler dans la cellule. Petit à petit, au rythme où la tension augmente, la part du courant passant dans la capacité diminue et celle passant dans la cellule augmente. Le temps caractéristique de cette montée de la tension est $R_{\text{cell}}C_p$, et la dérivée à l'origine (lorsque $V_{\text{cell}} = 0$) est $\frac{I_{\text{prog}}}{C_p}$ soit, par exemple, 1 V par 100 ns pour un courant de 100 µA et une capacité de 10 pF. La montée progressive de la tension rend l'écriture en courant plus semblable à une écriture utilisant une rampe de tension qu'à une écriture utilisant une source de tension constante. Comme l'utilisation d'une rampe de tension a été identifiée comme améliorant l'homogénéité du temps de SET ainsi que diminuant l'énergie requise [16], on pourrait s'attendre à obtenir des résultats similaires dans le cas de l'écriture en courant.

L'utilisation d'une source de courant présente également une autre particularité. Lorsque le SET se produit, la conductivité de l'élément mémoire OxRAM augmente rapidement. Cela a pour effet de faire tendre V_{cell} à diminuer, ce qui cause une décharge de la capacité. Dans la mesure où cette variation de

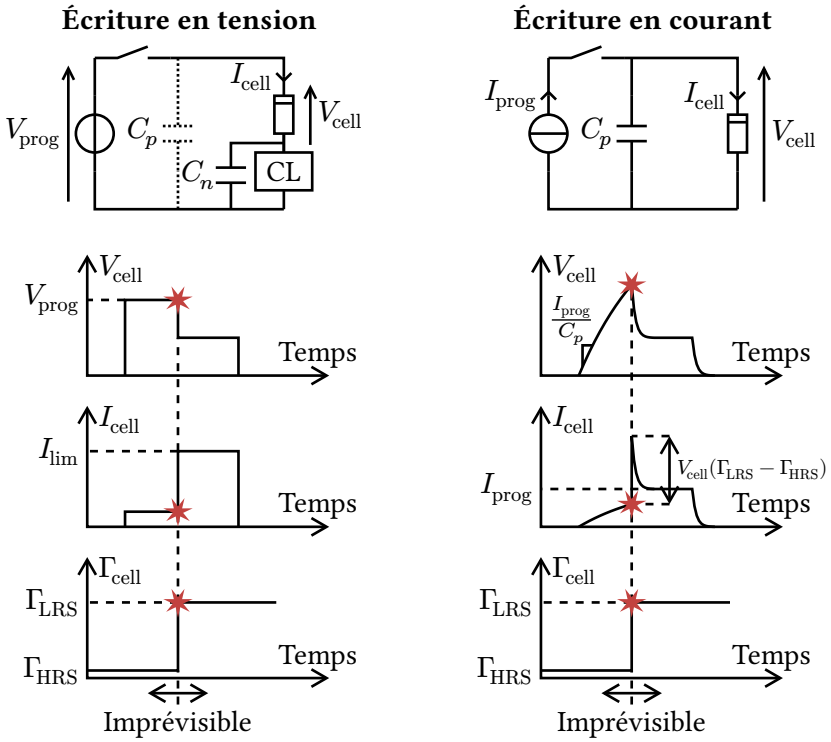


Figure 3.1 – Comparaison entre écriture en tension et écriture en courant en prenant en compte la capacité d'interconnexion C_p.

[25] Si l'on considérait une source de tension ayant une impédance de sortie R_s non nulle, C_p aurait un effet. Dans ce cas, elle limiterait la dynamique de montée de la tension. Le temps caractéristique serait alors $\frac{R_s R_{\text{cell}} C_p}{R_s + R_{\text{cell}}} \approx R_s C_p$ avec $R_s \ll R_{\text{cell}}$, ce qui est indépendant des grandeurs liées à la cellule OxRAM, contrairement au cas de la source de courant, même lorsqu'elle est parfaite.

conductivité est beaucoup plus rapide que la constante de temps électrique du circuit, cette décharge génère un pic de courant dans l'élément mémoire OxRAM. Si ce pic est assez grand, il a pour effet de dépasser temporairement la valeur de I_{prog}. Ce phénomène est similaire au phénomène de dépassement de courant observé dans le cas de la source de tension [51] [82]. Cependant, la capacité mise en jeu dans le cas de la source de tension n'est pas la capacité d'interconnexion C_p entre la source et l'élément mémoire OxRAM mais la capacité C_n du nœud reliant l'élément mémoire OxRAM à la limitation en courant (CL sur la Figure 3.1). Comme la source de courant est extérieure à la matrice, C_p est au moins aussi grande que C_n. L'effet dans le cas de la source de courant est donc au moins aussi important que dans le cas de la source de tension avec limitation de courant externe à la matrice.

Ce pic de courant présente l'opportunité d'écrire plus fortement l'élément mémoire OxRAM (I_c^{eff} plus grand donc Γ_{LRS} plus grand également) avec une énergie moindre. En effet, en comptant sur ce pic de courant, on pourrait arriver à un même Γ_{LRS} avec un courant de programmation I_{prog} plus faible que le courant de limitation I_{lim} de l'écriture en tension. Comme l'énergie consommée par le circuit intégré est directement proportionnelle au courant consommé (car l'alimentation du circuit est une source de tension constante), elle serait également réduite. La capacité servirait de stockage d'énergie et l'écriture serait hybride entre une écriture en courant et une écriture capacitive (l'écriture capacitive est discutée dans le Chapitre V).

En supposant la dynamique de la transition de la conductivité de l'élément mémoire OxRAM très grande devant le RC du circuit électrique, l'amplitude de la discontinuité sur I_{cell} est égale au produit entre la tension V_{cell} au moment de la transition et la différence de conductance $\Gamma_{\text{LRS}} - \Gamma_{\text{HRS}}$ entre les deux états. En particulier, elle est indépendante de la valeur de la capacité d'interconnexion C_p[26]. On a donc peu de contrôle sur ce pic qui dépend de grandeurs propres à l'élément mémoire OxRAM peu prédictibles. Si ce pic de courant affecte l'écriture de manière significative, il risque de réduire la prédictibilité et l'homogénéité de l'écriture. L'ajout d'une limitation de courant locale (proche de la cellule, de sorte que la capacité d'interconnexion du nœud entre la cellule et le limiteur soit faible) pourrait alors permettre de reprendre en partie le contrôle sur ces pics. Cependant le circuit *WT-PAOLA* ne permet pas d'ajouter une telle limitation.

Comme la tension lors du Forming est plus importante que lors d'un SET, et que la différence de conductivité entre l'état initial et le LRS l'est également, ce pic de courant est d'autant plus grand lors du Forming. Une attention particulière doit donc être apportée sur le Forming pour éviter de détruire l'élément mémoire OxRAM lors de la décharge capacitive.

III.3 Circuit intégré *WT-PAOLA*

III.3.1 Intentions initiales du circuit

Le circuit utilisé dans ce chapitre est le circuit *WT-PAOLA*. Il a été développé dans le cadre de la thèse de Paola Trotti [83], mais est arrivé trop tardivement pour pouvoir être testé. Ce circuit avait pour objet l'étude d'une technique d'écriture incorporant une méthode de « Write Terminaison ». L'idée était de ne pas directement utiliser une source de tension, mais plutôt d'utiliser une source de courant avec une régulation de la tension aux bornes de la source connectée à l'élément mémoire OxRAM. L'intérêt est de fixer la tension sur les temps longs tout en fixant le courant sur des temps courts. Ainsi durant le délai entre la transition SET et sa détection, le courant dans élément mémoire OxRAM serait maintenu par la source. La régulation, plus lente que ce délai, n'aurait pas le temps de réagir. Plutôt

[26]La hauteur maximale du pic de courant est égale au produit entre la tension à laquelle se produit le changement de conductance et la différence de conductance. Ce maximum n'est atteint que lorsque la dynamique du changement de conductance est bien supérieure à la dynamique de relaxation du couple RC. La capacité n'influence alors que l'énergie contenue dans ce pic en modulant le temps de relaxation de ce courant. Lorsque les dynamiques sont proches, le pic est « amorti » par la relaxation de couple RC avant d'atteindre son maximum [82].

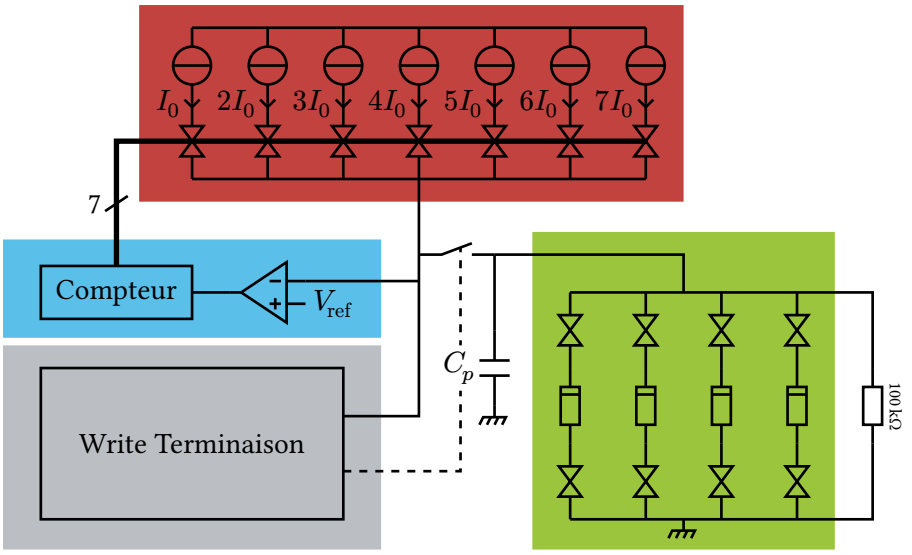


Figure 3.2 – Schéma simplifié du circuit *WT-PAOLA*.

que d'avoir besoin d'une limitation de courant durant ce délai, la tension chuterait d'elle même. Cela permettrait donc d'avoir une fin d'écriture rapide, sans investir dans un circuit de détection rapide. La détection de la transition se fait alors en observant cette chute de tension, et non pas l'augmentation du courant, comme c'est le cas dans les systèmes utilisant une source de tension. Après la détection, l'écriture peut être arrêtée avant que la régulation n'ait eu le temps de réagir.

L'intention derrière le développement du circuit *WT-PAOLA* est donc bien différente de ce que nous souhaitons en faire. Nous serons donc inévitablement confrontés à des limites imposées par l'architecture du circuit. Néanmoins, cela représente une bonne opportunité de réaliser une première expérimentation de l'écriture en courant avant de lancer le développement d'un circuit spécial comme celui présenté dans le Chapitre IV.

III.3.2 Structure

La Figure 3.2 illustre la structure du circuit *WT-PAOLA*. Il est composé de :

- Une source de courant programmable sur 7 bits dont seuls les 3 bits de poids forts sont éventuellement accessibles depuis l'extérieur
- Un compteur 7 bits commandant la source de courant
- Un comparateur comparant la tension sur l'élément mémoire OxRAM avec une tension cible permettant de faire converger la tension sur l'élément mémoire OxRAM vers la tension cible par rétroaction en contrôlant la direction de comptage

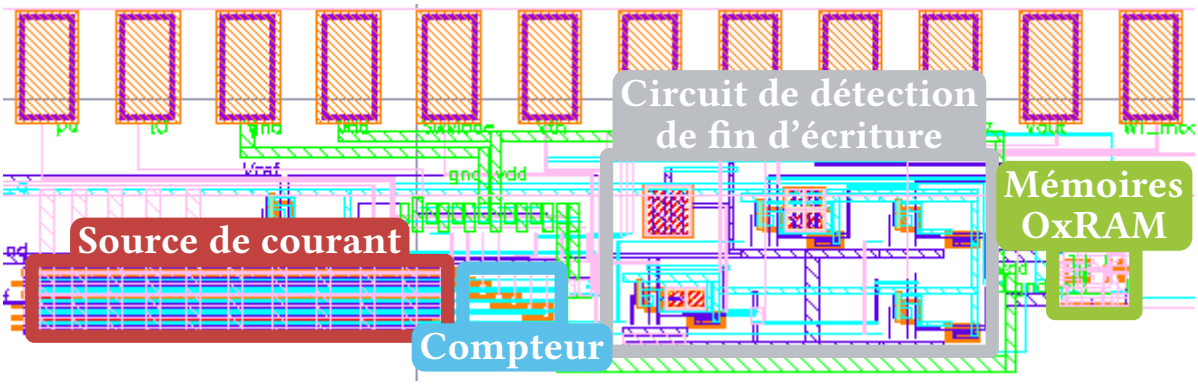


Figure 3.3 – Dessin des masques (« layout » en anglais) du circuit *WT-PAOLA*.

- 4 éléments mémoires OxRAM sélectionnables au travers de portes de transmission, ce ne sont pas des cellules 1T1R, il n'y a pas de transistor en série, et le courant ne peut donc pas être limité
- Un circuit de détection de fin d'écriture par observation de la tension sur l'élément mémoire OxRAM
- Une résistance de 100 kΩ en parallèle de l'élément mémoire OxRAM pour limiter la tension

La Figure 3.3 montre le dessin des masques (« layout » en anglais) du circuit *WT-PAOLA*.

III.3.3 Utilisation effective du circuit

Dans notre cas, nous souhaitons étudier l'écriture en courant. Nous sommes donc intéressés par la source de courant programmable intégrée. Elle nous permettra de jouer le rôle de la source de courant. Par contre, nous ne souhaitons pas utiliser la régulation de la tension, ni la détection du SET. Un ensemble de mode de débogage permet d'obtenir ce comportement, ce qui nous permet d'obtenir la configuration que l'on souhaite. Cependant, du fait d'un défaut de normalisation de l'horloge du compteur, ce dernier ne peut pas être utilisé de manière fiable pour fixer les 7 bits de la source en utilisant un nombre de coups d'horloge prédéfini. Nous avons donc dû nous limiter aux 3 bits de poids forts directement accessibles depuis l'extérieur du circuit. La capacité d'interconnexion notée C_p dans les parties précédentes a été estimée à 700 fF par des simulations prenant en compte le dessin des masques du circuit.

Les tests ont été réalisés sous pointes sur des galettes entières. Les différentes tensions et courants fixés/observés le sont au travers d'un SMU (B1500). La partie numérique est contrôlée par un arduino. Une interface de test a été développée en python pour l'instrumentation logicielle (commande de l'arduino, des SMU et des autres instruments de laboratoire).

III.4 Résultats expérimentaux

III.4.1 Forming

Dans le circuit *WT-PAOLA*, les éléments mémoires OxRAM sont seuls, sans transistor en série permettant de limiter le courant. On ne peut donc pas limiter le courant dans l'élément mémoire OxRAM. Bien que l'utilisation d'une source de courant[27] devrait permettre de se passer d'une limitation du courant, puisque ce dernier devrait être fixé par la source, nous avons vu que la décharge de la capacité d'interconnexion C_p pouvait engendrer un pic de courant. Pendant le Forming, ce pic de courant est particulièrement grand à cause de la tension de Forming élevée et de la grande différence de conductance entre l'état initial et l'état écrit. Ce pic a la capacité de détruire l'élément mémoire OxRAM en rendant la réaction essentiellement irréversible. Nous avons donc rencontré des difficultés pour former les éléments mémoires OxRAM sur ce circuit.

Pour maximiser nos chances de réussite lors de l'étape de Forming d'un élément mémoire OxRAM, nous avons procédé à un Forming progressif pour essayer de répartir la différence de conductance à traverser en plusieurs fois, et donc en plusieurs pics de courant d'amplitude plus faible. La Figure 3.4 illustre la procédure utilisée. Nous procédons en effectuant une séquence d'impulsion de courant. Nous commençons par l'amplitude I_{FORM} la plus faible disponible. Pendant chaque impulsion, la tension résultante sur l'élément mémoire OxRAM V_{FORM} est observée à l'aide d'un oscilloscope. La conductance Γ de l'élément mémoire OxRAM est ensuite mesurée de manière standard (en appliquant une faible tension de 0,2 V et en mesurant le courant). Si Γ est supérieur à une valeur prédéterminée Γ_{cible}, le Forming est terminé et cette procédure prend fin. Dans le cas contraire, l'amplitude du courant est augmentée progressivement jusqu'à atteindre une tension sur l'élément mémoire OxRAM cible V_{cible}. Lorsque cette tension est atteinte, l'impulsion de courant est répétée sur l'élément mémoire OxRAM à

[27] Nous n'avons pas considéré l'utilisation d'une source de tension pour l'écriture ou le Forming des éléments mémoires OxRAM du circuit *WT-PAOLA* en raison de l'absence de limitation en courant intégrée. L'ajout d'une limitation externe présenterait les mêmes problématiques de décharge capacitive que la source de courant intégrée.

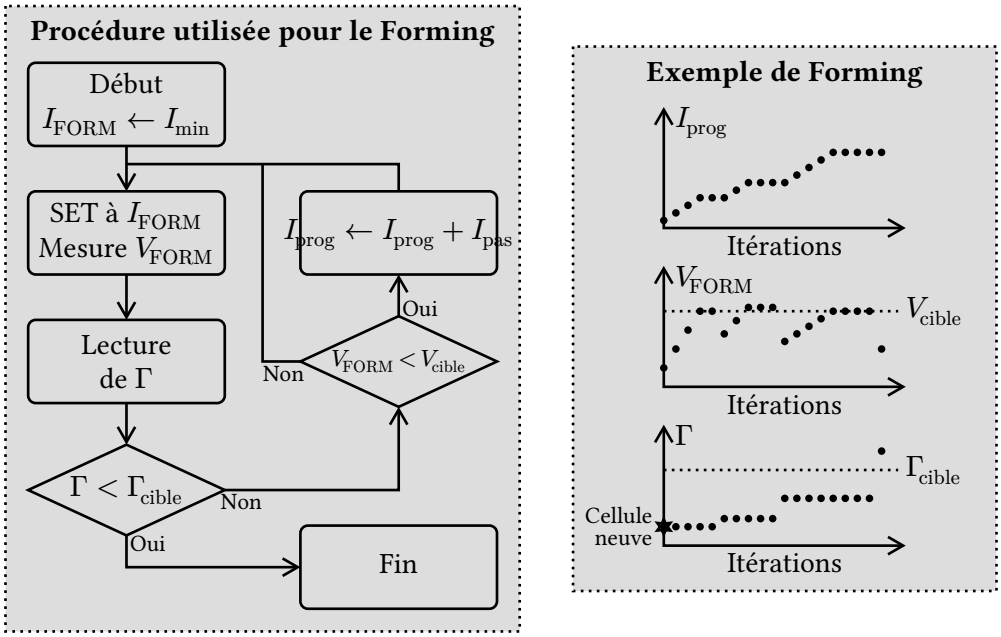


Figure 3.4 – Description de la procédure de formation des filaments utilisée (gauche). Diagramme de la procédure (droite). Exemple d'exécution de cette procédure.

l'identique jusqu'à ce que Γ augmente significativement. Éventuellement, Γ sera assez haut pour que la procédure prenne fin. Cette procédure est effectuée manuellement et n'est pas infaillible. Certains éléments mémoires OxRAM sont tout de même détruits lors de leur Forming, d'autres semblant bien se comporter lors de cette procédure semblaient dysfonctionnels lors de leur utilisation, etc… Nous avons donc dû éliminer un certain nombre d'éléments mémoires OxRAM en considérant que leur Forming avait échoué. Cette élimination présente un risque élevé de biais de sélection (« cherry-picking » en anglais) car il est difficile de statuer objectivement si le Forming s'était bien passé ou non. De plus, le fait que le Forming de chaque élément mémoire OxRAM soit différent peut également cacher des biais. En effet, l'historique des états de l'élément mémoire OxRAM, en particulier ceux du Forming par exemple, pourrait jouer un rôle plus ou moins important dans son comportement futur. Après tout l'élément mémoire OxRAM est une mémoire.

Comme l'étape de Forming des éléments mémoires OxRAM du circuit *WT-PAOLA* était difficile et longue, nous n'avons pas pu en former avec succès un grand nombre, environ une dizaine. Nous avons donc dû effectuer la suite de nos expériences sur un nombre limité d'éléments mémoires OxRAM. Ainsi, il sera difficile dans ce chapitre de juger de l'homogénéité ou de la répétabilité de l'expérience. Ces difficultés proviennent de l'utilisation détournée que nous faisons du circuit *WT-PAOLA*. Elles seront réglées dans le Chapitre IV qui considéra un circuit spécialement développé pour ces expériences. Les observations dans ce chapitre se concentreront donc sur des dispositifs unitaires et sur des statistiques purement cycle-à-cycle.

III.4.2 Effet de l'amplitude du courant

Après avoir réussi à former un élément mémoire OxRAM, nous devrions pouvoir l'écrire et l'effacer. Ici, nous nous concentrerons sur l'écriture ou le SET car c'est pour lui que l'utilisation de la source de courant semble *a priori* intéressante. Pour le RESET, nous utiliserons une source de tension externe. L'absence de limitation en courant ne pose pas de problème en RESET puisque la conductance de l'élément mémoire OxRAM diminue. Le choix des paramètres du RESET a été fait au travers de l'*a priori* issu de l'expérience du laboratoire sur ces dispositifs et d'expériences préliminaires non décrites ici. Elle consistait a simplement vérifier que le RESET était convenablement effectué avec ces paramètres.

Nous utiliserons donc dans la suite un RESET à 2 V et durant 10 µs (sauf dans la première expérience qui a une impulsion de 100 µs).

Le SET est quant à lui toujours effectué avec la source de courant intégré. L'absence de limitation en courant nous empêche d'effectuer un SET avec une source de tension externe, puisque même si elle était elle-même limitée à l'extérieur du circuit, la capacité d'interconnexion entre l'élément mémoire OxRAM et le limiteur provoquerait des pics de courant dans l'élément mémoire OxRAM. Nous ne pourrons donc pas directement comparer un SET en courant avec un SET en tension dans cette expérience. Le circuit développé et exploité dans le Chapitre IV permettra une telle comparaison.

Le premier paramètre que nous étudions est celui de l'amplitude de l'impulsion de courant lors du SET. Comme nous l'avons vu dans l'introduction de cette partie sur l'écriture des OxRAM, la réaction constituant le SET est principalement gouvernée par le champ électrique dans l'oxyde. Ainsi, si le produit du courant et de la résistance en HRS (état avant le SET) est trop faible, la tension ne pourra pas être suffisamment élevée pour provoquer une réaction dans l'oxyde dans la durée de l'impulsion. On s'attend donc à constater un minimum d'amplitude de l'impulsion en courant nécessaire. De même, en ignorant le pic de courant pour l'instant, plus l'amplitude de l'impulsion en courant est grande, plus la résistance de l'élément mémoire OxRAM en LRS devra être faible pour que la tension baisse suffisamment pour arrêter la réaction. Si le pic de courant ne joue aucun rôle dans le SET, nous devrions donc avoir une conductance en LRS positivement corrélée avec l'amplitude de l'impulsion de courant, comme elle l'est avec la valeur du courant de limitation lors d'une écriture en tension. Sinon, si le pic de courant joue un rôle important, nous pourrions observer des niveaux de conductance en LRS supérieurs à ceux attendus avec un tel courant d'écriture, et la corrélation pourrait être moins forte puisque l'amplitude de l'impulsion de courant ne joue pas directement sur l'amplitude du pic, mais indirectement au travers de la vitesse de montée de la tension sur l'élément mémoire OxRAM, et donc sur la tension de SET effective $V_{\mathrm{SET}}^{\mathrm{eff}}$.

Pour savoir dans quelle situation nous sommes, nous avons effectué l'expérience sur un de nos éléments mémoires OxRAM correctement formé. Nous avons réalisé 1 000 cycles consécutifs avec une impulsion de courant durant 100 µs à une certaine amplitude, et un RESET de 100 µs à 2 V. Nous

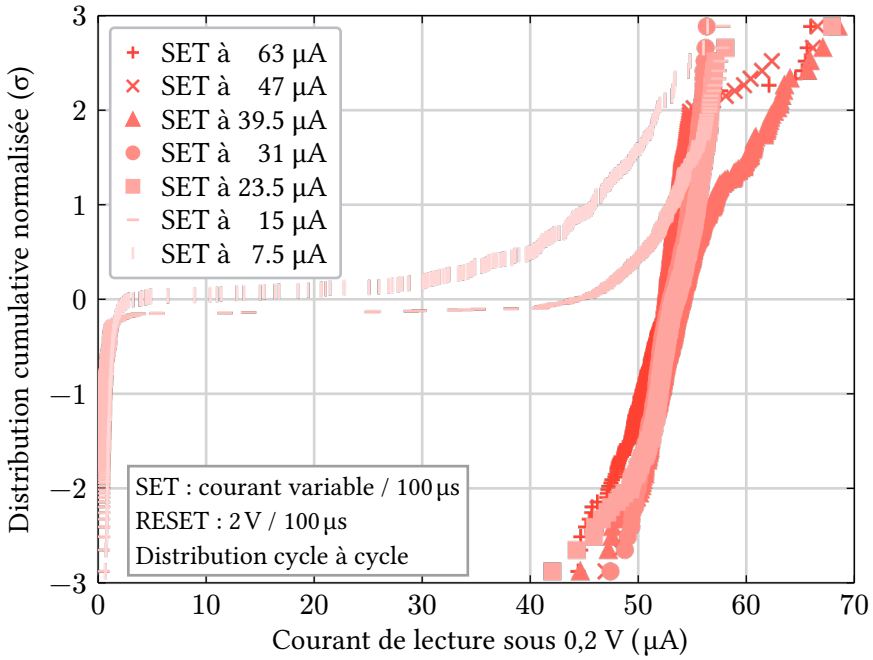


Figure 3.5 – Distribution cumulative normalisée du courant de lecture après un SET en courant pour différentes amplitudes d'impulsion.

utilisons un temps de SET relativement long pour être sûrs que la tension sur l'élément mémoire OxRAM atteigne bien sa valeur finale même lorsque le courant est faible. Ces 1 000 cycles sont répétés pour différentes amplitudes de l'impulsion en courant. Les distributions des 500 derniers cycles de chaque groupe de 1 000 cycles sont représentées sur la Figure 3.5. Les 500 premiers cycles ne sont pas considérés pour diminuer les biais pouvant apparaître à cause du changement de condition et de l'effet mémoire. L'expérience est réalisée de l'amplitude la plus forte à la plus faible.

On constate que les distributions ainsi obtenues sont similaires pour toutes les amplitudes supérieures ou égales à 23,5 µA. En-dessous de cette amplitude, l'élément mémoire OxRAM n'est pas correctement écrit pour plus de la moitié des cycles. Nous nous attendions à constater un tel seuil. Nous pouvons vérifier qu'il se trouve bien où l'on pouvait l'attendre. En HRS, notre courant de lecture médian est environ 5 µA. La lecture est effectuée sous 0,2 V. Si nous estimons qu'une tension supérieure à 0,8 V est nécessaire pour que l'élément mémoire OxRAM s'écrive de façon assez certaine, un courant d'au moins 20 µA est nécessaire (sans prendre en compte la non-linéarité de la conductance en HRS qui aurait tendance à augmenter ce courant seuil). Cela s'accorde donc avec notre observation qui place ce seuil entre 15 µA et 23,5 µA.

Par contre, nous ne constatons pas de corrélation entre l'amplitude de l'impulsion de courant, lorsqu'elle est suffisante pour SET l'élément mémoire OxRAM, et la conductivité en LRS. Cela indiquerait donc que le pic en courant joue un rôle important dans l'écriture. Pour vérifier cela, nous estimons le courant limite I_c^{eff} équivalent nécessaire à l'écriture des éléments mémoires OxRAM à la conductance observée avec une source de tension. La médiane de courant de lecture en LRS est légèrement supérieure à 50 µA, mais prenons ce nombre pour simplifier la suite. La lecture est effectuée à 0,2 V, et on estime que la tension d'équilibre de la réaction est 0,7 V (celle qui tend à être atteinte par la réaction dans l'oxyde). Le courant limite nécessaire pour atteindre cet état de conductance, qui est le courant que l'on aurait dans cet état lorsque la réaction serait à l'équilibre est donc 175 µA. Ce nombre est bien supérieur au courant que notre source de courant n'aurait jamais fourni. Donc il semble bien que le pic de courant affecte le SET de manière importante.

Avec une écriture aussi importante, même avec la plus grande amplitude de courant que nous avons utilisée, 63 µA, la tension sur l'élément mémoire OxRAM après le SET, et après le pic de courant, ne serait que de 0,25 V, soit environ la tension de lecture. L'écriture serait donc totalement arrêtée après le pic de courant. Cela explique donc pourquoi on observe que l'amplitude de l'impulsion de courant n'a pas d'effet au-delà du seuil précédemment discuté. De plus, cela confirme également que l'amplitude de l'impulsion de courant influence peu l'amplitude du pic de courant. La combinaison du pic de courant ainsi que l'utilisation d'une source de courant agissent donc comme une terminaison d'écriture.

En somme, la décharge capacitive est un élément essentiel de cette écriture. Elle permet d'écrire un élément mémoire OxRAM à une certaine conductance avec un courant moyen plus faible que le courant de limitation qui serait nécessaire dans le cas de l'utilisation d'une source de tension. Cela permet donc d'écrire un élément mémoire OxRAM à un même état avec une énergie moindre. Comme l'amplitude de l'impulsion de courant a peu d'importance, il est plus intéressant de le choisir faible, tout en gardant une marge avec le courant seuil établi précédemment. Cependant, aucun paramètre directement contrôlable n'a été identifié comme ayant une influence sur le niveau de conductivité atteint par l'élément mémoire OxRAM. Cela limite donc son utilisation à la programmation sur seulement deux niveaux de conductances, ce qui fait qu'une écriture multi-niveaux est impossible.

III.4.3 Effet de la durée de l'impulsion

On s'interroge maintenant sur l'effet de la durée de l'impulsion de courant. A priori, une impulsion trop courte ne laissera pas le temps à la tension sur l'élément mémoire OxRAM de monter. On s'attend donc à pouvoir trouver une durée minimale en-dessous de laquelle l'écriture ne se produit pas, comme pour

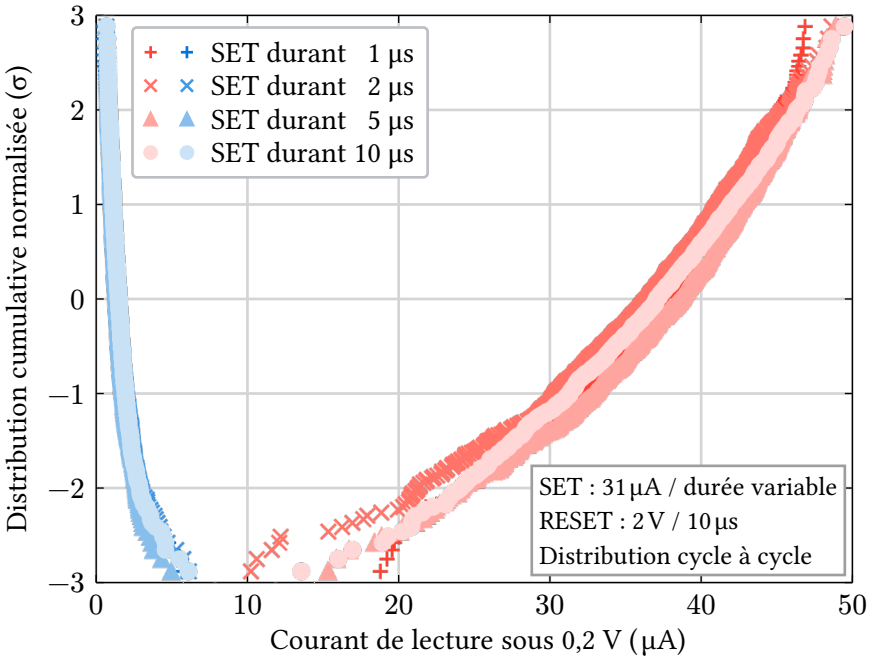


Figure 3.6 – Distribution cumulative normalisée du courant de lecture après un SET (rouge) et après un RESET (bleu) pour différentes longueurs d'impulsion de SET.

l'amplitude. Par contre, on ne s'attend pas à observer une grande influence de la durée de l'impulsion une fois au-dessus de ce seuil. En effet, comme nous venons de le voir, après le SET, la tension sur l'élément mémoire OxRAM est de l'ordre de la tension de lecture, c'est-à-dire une tension qui n'est pas assez grande pour engendrer un changement de l'état de l'élément mémoire OxRAM. Comme l'instant où le SET se produit est indépendant de la durée de l'impulsion, augmenter la durée de l'impulsion revient à augmenter le temps passé dans cette situation. Une impulsion longue ne devrait donc pas être différente d'une plus courte, à la différence d'une légère tension rémanente équivalente à une lecture. Il y a un effet d'auto-terminaison de l'écriture, ce qui était l'effet recherché par ce circuit lors de son développement.

Pour vérifier cela, on effectue sur un second élément mémoire OxRAM la même procédure que pour tester l'effet de l'amplitude, mais en faisant varier la durée de l'impulsion à la place, avec une amplitude fixée à 31 µA. La Figure 3.6 montre les distributions ainsi obtenues. On constate que les quatre distributions sont très similaires. Les durées plus grandes semblent donc bien équivalentes à celles plus faibles. Cependant, nous n'identifions pas de durée seuil. La durée minimale testée est 1 µs, ce choix n'est pas arbitraire, il correspond à la durée d'impulsion la plus courte que nous pouvions contrôler avec notre banc de test. On suppose donc que ce seuil existe bien mais que nous n'avons pas pu l'observer à cause de cette limite expérimentale. Cependant, on peut noter que ce seuil est nécessairement inférieur à 1 µs, qui n'est pas une durée particulièrement élevée pour l'écriture des OxRAM. Ainsi notre méthode d'écriture, qui pourrait sembler *a priori* limitée en vitesse par la charge de la capacité ne l'est pas vraiment en pratique. Si l'on estime que C_p vaut 700 fF, avec un courant de 31 µA, la dérivée de la tension au début de l'écriture est de plus de 40 V/us et n'est donc pas l'effet limitant. Pour s'assurer que cette écriture n'est pas limitée en terme de vitesse à comparer les autres méthodes d'écriture, il aurait été préférable de tester une durée de 100 ns, durée souvent ciblée par les implémentations, mais nous n'avons pas pu le faire.

Le lecteur avisé remarquera également que les distributions LRS présentées par la Figure 3.5 et la Figure 3.6 n'ont pas du tout la même variance bien qu'elles aient été obtenues dans des conditions similaires. La principale différence entre ces deux expériences est l'élément mémoire OxRAM utilisé.

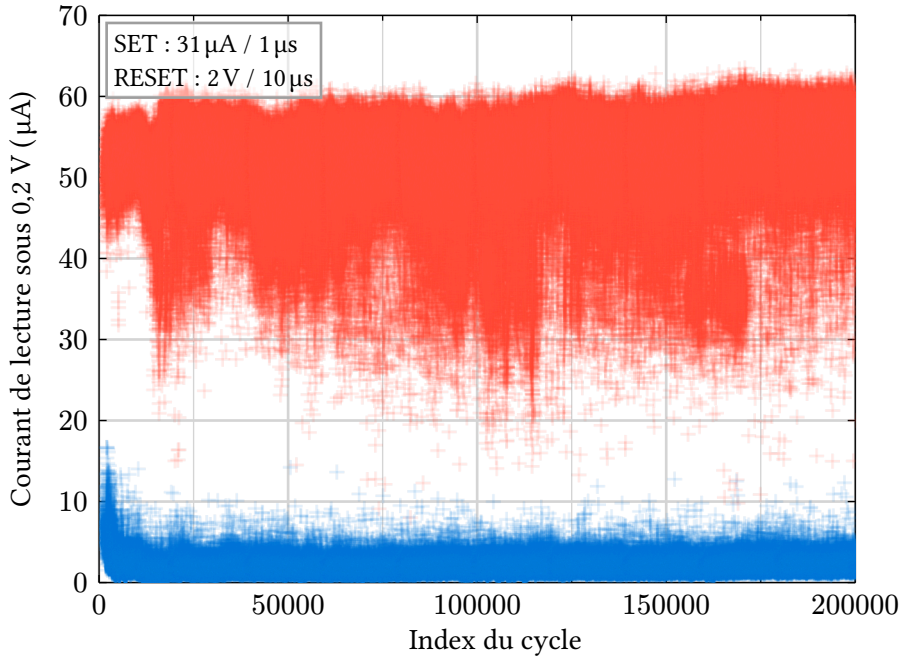


Figure 3.7 – Nuage de points de la conductance de l'élément mémoire OxRAM au travers de 200 000 cycles, mesurée après le (rouge) SET, (bleu) RESET.

Nous supposons donc que cette différence provient d'un Forming différent. En effet, comme expliqué précédemment, le Forming est assez mal contrôlé et incertain à cause d'une contrainte liée au circuit utilisé, en particulier d'un manque d'une limitation en courant intégré au circuit.

III.4.4 Vérification de l'endurance

Nous avons ensuite souhaité voir comment cette méthode influençait l'endurance de l'élément mémoire OxRAM. D'un côté, les niveaux de conductance obtenue en LRS sont relativement élevés, or l'endurance des éléments mémoires OxRAM a été observée comme corrélée négativement avec la valeur de conductance LRS [84]. On pourrait donc s'attendre à observer un vieillissement rapide. D'un autre côté, notre méthode arrête l'excitation très rapidement après le SET, comme le font les méthodes de « Write Terminaison ». Cependant, ces méthodes promettent une augmentation de l'endurance [25]. On pourrait par conséquent s'attendre à obtenir une bonne endurance.

Pour voir ce qu'il en est expérimentalement, nous avons, sur un troisième élément mémoire OxRAM, effectué 200 000 cycles avec une SET en courant d'amplitude 31 µA durant 1 µs et un RESET durant 10 µs sous 2 V. La Figure 3.7 présente la conductance aussi bien HRS que LRS lue après chaque opération d'écriture/effacement effectuée sous forme de nuage de points avec en ordonné le courant mesuré sous 0,2 V et en abscisse l'index de la mesure. Pendant les 2 000 premiers cycles, la moyenne de HRS est plus grande que dans la suite. Ces premiers cycles, où la distribution HRS évolue rapidement est connue et souvent dénommée phase de « seasoning ». Généralement après le Forming d'une cellule, un certain nombre de cycles sont réalisés pour passer cette phase et étudier la phase suivante qui présente des caractéristiques plus stables.

Ces 200 000 cycles ne sont pas suffisants pour observer la limite d'endurance, là où la fenêtre se refermerait, mais ils permettent de s'assurer qu'au moins 200 000 cycles sont réalisables.

Les données LRS sont, pendant les 10 000 premiers cycles, distribuées avec une variance faible, semblable à celle observée sur la Figure 3.5 lors de l'identification du rôle que joue l'amplitude de l'impulsion de courant. Après ces 10 000 premiers cycles, la variance augmente subitement vers quelque chose ressemblant plus à ce qui est observé sur la Figure 3.6 lors de l'étude de l'effet de

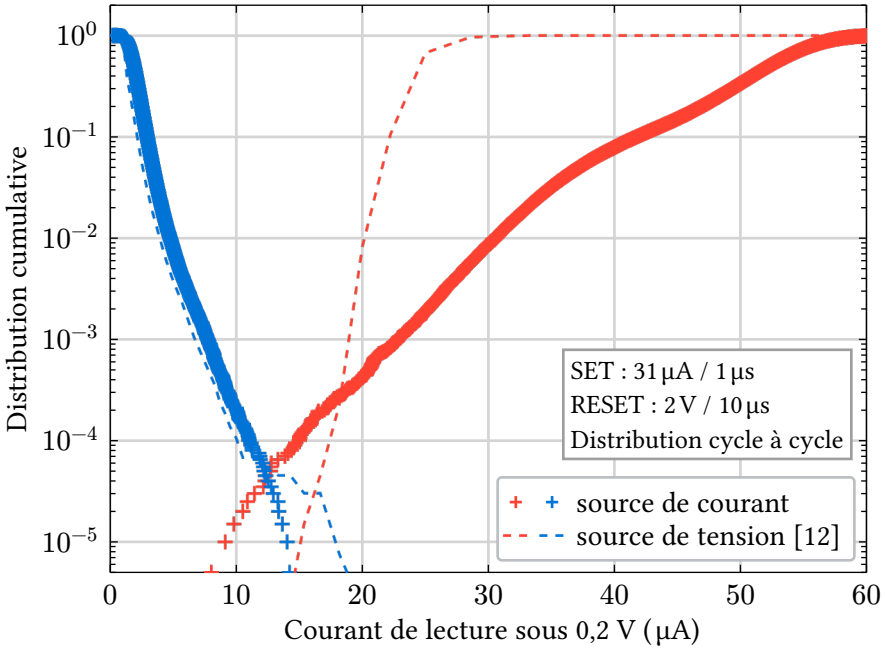


Figure 3.8 – Comparaison de la distribution obtenue pendant le cyclage avec l'écriture en courant (mêmes données que sur la Figure 3.7 exceptés les 2 000 premiers cycles) et d'un exemple de distribution obtenue par écriture en tension (extrait de [12])

la durée de l'impulsion. On voit donc que différentes gammes de variance sont possibles. Toutefois, la raison du passage d'un mode à l'autre n'est pas claire. On pourrait penser d'après l'observation de ces 200 000 cycles que l'augmentation de la variance en LRS est liée au vieillissement. Mais cela n'expliquerait pas le cas des données de la Figure 3.6 qui ont été obtenues dans les 5 000 premiers cycles de la vie de l'élément mémoire OxRAM utilisé, alors que les données de la Figure 3.5 présentant une variance faible ont été obtenues lors des 7 000 premiers cycles. La distribution de la conductance en LRS semble également varier au travers du temps, avec des moments où la variance est plus grande qu'à d'autres.

La Figure 3.8 montre les mêmes données que la Figure 3.7 (exceptés les 2 000 premiers cycles dit de « seasoning ») sous forme de distribution cumulative. On peut y mesurer un taux d'erreur à marge de lecture nulle (supposant donc un circuit de lecture parfait, et un bruit de lecture nul) de $5.10^{-5.}$ Une distribution obtenue sur la même technologie d'élément mémoire OxRAM, mais utilisant une écriture en tension est également rapportée pour comparaison. Elle provient de [12]. Un taux d'erreur similaire est obtenu. Cependant, on peut constater que la distribution LRS est beaucoup plus large avec l'écriture en courant utilisée ici. En particulier, la médiane de conductance en LRS est beaucoup plus élevée avec l'écriture en courant qu'avec l'écriture en tension.

III.4.5 Écriture avec vérification

Le taux d'erreur de 5.10^{-5} n'est pas satisfaisant pour réaliser une mémoire qui contient des millions d'éléments mémoires OxRAM. Pour palier à cela, une solution communément utilisée, et déjà discutée dans le Chapitre II.3.4, est la vérification. Ici nous considérons la méthode que nous avions nommée la ré-écriture avec vérification. Elle consiste à effectuer une lecture directement après le SET pour vérifier que le niveau de conductance atteint soit convenable et, s'il ne l'est pas, l'élément mémoire OxRAM est d'abord effacé, puis ré-écrit. On peut procéder de la même manière pour le RESET de la cellule. Si la conductance résultante n'est pas assez faible, l'élément mémoire OxRAM sera d'abord écrit avant d'être ré-effacé.

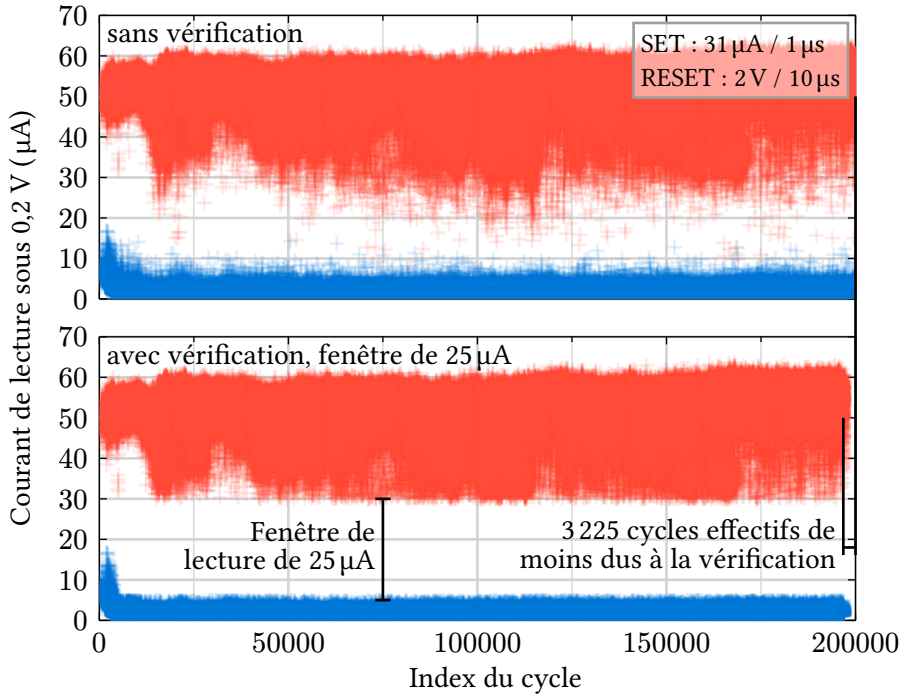


Figure 3.9 – Comparaison entre (au-dessus) le nuage de points de la Figure 3.7 et (en-dessous) le même nuage de points donc les points marqués virtuels par l'algorithme de vérification ont été retirés. La fenêtre de lecture ciblée par l'agorithme est 25 µA. L'agorithme n'est pas activé pendant les 2 000 premiers cycles correspondants au « seasoning ».

Ici, plutôt que d'implémenter cette méthode sur notre banc de test, et de refaire de nouvelles mesures, nous quantifions l'effet de cette procédure sur notre écriture en la simulant sur les données issues du

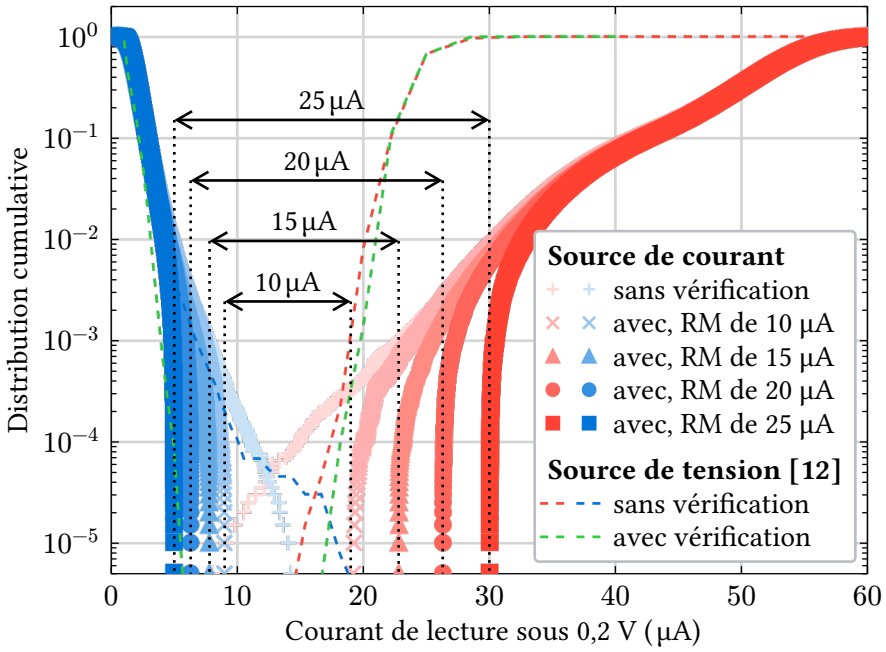


Figure 3.10 – Comparaison entre les distributions de la Figure 3.8 et les distributions correspondantes après avoir appliqué l'algorithme de vérification pour différentes fenêtres de lecture (RM) cibles 10 µA, 15 µA, 20 µA et 25 µA. La distribution obtenue par écriture en tension (extrait de [12]), ainsi que sa distribution après vérification sont également rapportées pour comparaison.

test d'endurance (Figure 3.7 et Figure 3.8). Pour ce faire, on fixe deux seuils de conductance, un pour le SET et un pour le RESET. Puis pour chaque point mesuré, on vérifie s'il est inférieur au seuil défini pour le RESET s'il s'agit d'une lecture après un RESET, ou s'il est supérieur au seuil défini pour le SET s'il s'agit d'une lecture après un SET. Si un point mesuré ne vérifie pas cette condition, ce point, ainsi que le suivant, sont marqués comme virtuels, cela représente la partie « retente notre chance » de l'algorithme. Ces points virtuels correspondent exactement à ceux qui auraient été cachés par l'algorithme de vérification. La Figure 3.9 montre l'effet de cette procédure sur le nuage de points de la Figure 3.7 en omettant les points marqués virtuels avec des seuils choisis pour obtenir une marge de lecture de 25 µA. On voit clairement apparaître une zone entre les points LRS et les points HRS où plus aucun point n'apparaît car ils ont été rejetés par l'algorithme de vérification. On constate également que le nombre de cycles utiles a diminué puisque 3 225 couples de points ont été marqués virtuels car ils ne respectaient pas les seuils imposés.

La Figure 3.10 illustre les distributions obtenues en suivant cette procédure pour différentes valeurs cibles de largeur de fenêtre de lecture. Les seuils sont choisis de sorte à minimiser le nombre de points rejetés tout en assurant la fenêtre demandée. En pratique, cela revient à chercher deux seuils séparés par la fenêtre de lecture souhaitée et tels qu'ils intersectent leur distribution respective à la même ordonnée. Cette ordonnée correspond d'ailleurs à la probabilité qu'une écriture ne respecte pas le seuil lui étant associé. L'application de cette méthode de choix des seuils requiert bien sûr d'avoir un *a priori* sur ces distributions. Dans notre cas, l'*a priori* n'est pas nécessaire puisque notre simulation est faite *a posteriori* en connaissance de toutes les données.

Comme la variance de la distributions LRS est plus grande que celle de la distribution HRS, c'est le seuil LRS qui bouge le plus entre les différentes valeurs de largeur de fenêtre cible. Sur cette même figure les distributions à la fois avec et sans vérification présentées dans [12], une implémentation utilisant une écriture en tension, sont rapportées pour comparer. On constate que dans le cas de l'écriture en tension, la vérification permet principalement de réduire le support de la distribution HRS pendant que la distribution LRS n'est presque pas affectée. Alors que dans notre cas, la majorité de la fenêtre de lecture est gagnée sur le territoire de la distribution LRS. Ainsi, la large variance de la distribution LRS

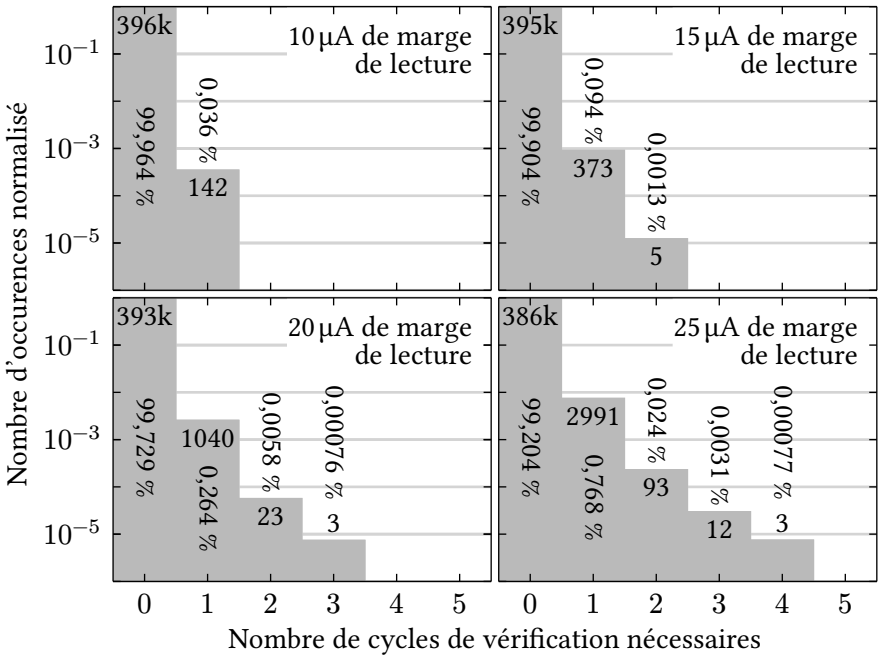


Figure 3.11 – Histogrammes représentant le nombre de cycles ayant nécessité un certain nombre de cycles de vérification pour différentes marges de lecture cibles.

avec l'écriture en courant, qui semble *a priori* être un inconvénient, est, dans le cas de la vérification, un avantage permettant d'obtenir de plus larges fenêtres de lecture pour un même taux de rejet.

La vérification à un coup, en rejetant des écritures, elle accélère le vieillissement en cyclant l'élément mémoire OxRAM plus souvent que l'utilisateur ne le voit. L'énergie requise est également augmentée par la multitude d'écriture. Enfin, la durée d'une écriture devient aléatoire et possiblement plusieurs fois plus longue qu'une écriture standard, car de nombreuses tentatives peuvent avoir lieu. Ces inconvénients sont d'autant plus importants que le nombre de rejets est grand. La Figure 3.11 montre la répartition du nombre de cycles nécessaires dans notre simulation pour satisfaire le critère de vérification pour les différentes valeurs de largeur de fenêtre de lecture ciblées. Plus la fenêtre ciblée est large, plus le nombre de cycles de vérification nécessaires augmente, car plus une large partie de la distribution naturelle doit être rejetée, ce qui augmente la probabilité d'occurrence.

Si les cycles étaient parfaitement indépendants entre eux, on s'attendrait à obtenir une progression exponentielle du nombre d'occurrences où au moins x cycles ont été nécessaires[28] en fonction de x. Néanmoins, on obtient légèrement plus d'occurrences avec x élevés que prédit par ce modèle, donc, il semblerait qu'il y ait une légère corrélation inter-cycles favorisant les échecs consécutifs. Une corrélation entre cycles à portée limitée avait déjà été observée par [85].

Dans tous les cas, le nombre de cycles supplémentaires effectués est faible, seulement 0,84 % avec notre largeur de fenêtre la plus agressive de 25 µA. Ainsi, les performances moyennes de la mémoire ne seraient pas négativement affectées par cette procédure. Il est intéressant d'observer que l'obtention d'une telle largeur de fenêtre sur les données issues de l'écriture en tension [12] présentées sur la Figure 3.10 nécessiterait de rejeter plus de la moitié des points, contre 1 sur 100 avec notre méthode d'écriture, ce qui aurait un impact non négligeable sur les performances moyennes. Ainsi, bien que la distribution LRS de l'écriture en courant semble moins bonne que celle de l'écriture en tension, elle est plus facilement et plus efficacement corrigeable par un algorithme de vérification simple tel que celui simulé ici.

III.5 Conclusion

L'utilisation d'un circuit préexistant (*WT-PAOLA*) nous a permis de pouvoir expérimenter notre idée d'écriture en courant. Cependant, elle s'est aussi avérée être un obstacle pour certains aspects. Les premiers résultats obtenus indiquent que la capacité d'interconnexion C_p joue un rôle important dans l'écriture en courant.

Nous avons observé que l'amplitude du courant joue finalement peu sur la distribution LRS, il est uniquement nécessaire que l'amplitude soit supérieure à un seuil correspondant au courant nécessaire pour provoquer une tension sur l'élément mémoire OxRAM supérieure à sa tension seuil d'écriture. De plus, la longueur d'impulsion joue également peu sur la distribution LRS, ce qui est la conséquence de l'effet d'auto-terminaison de ce type d'écriture, même si l'impulsion continue après le SET, la tension est suffisamment faible pour ne plus affecter l'élément mémoire OxRAM. Bien que nous n'ayons pas pu l'observer, un durée d'impulsion minimale, dépendant de l'amplitude en courant et de la valeur de la capacité d'interconnexion, devrait être nécessaire. Nous avons pu montrer qu'elle était inférieure à 1 µs dans le cas du circuit *WT-PAOLA*, ce qui est raisonnablement rapide. Cette manière d'écrire ne semble pas avoir affecté l'endurance de façon négative, 200 000 cycles ont été effectués sans observer de dégradations significatives du comportement. La distribution LRS a tendance à avoir une variance plus forte qu'avec une écriture en tension, en raison probablement du peu de contrôle que nous avons sur

[28] Si la probabilité de réussir un SET, c'est-à-dire, la probabilité d'écrire l'élément mémoire OxRAM a une conductance supérieure au seuil défini, est indépendante des précédentes tentatives et identique à celle des autres tentatives, elle correspond alors à une épreuve de Bernoulli. Le nombre d'échecs avant un succès suit alors, par définition, une loi géométrique dont la fonction de masse prend la forme $(1-p)^k p$ avec p la probabilité de réussite et k le nombre d'échecs avant le premier succès.

le pic de courant qui a été identifié comme le principal élément agissant sur le niveau de conductance LRS. Cependant, cette large variance, avec une médiane décalée vers les conductances plus élevées, permet d'obtenir de plus larges fenêtres de lecture à l'aide de la vérification, comparée à celle qu'on peut obtenir avec une écriture en tension.

Il est cependant nécessaire de modérer ces conclusions. En effet, les difficultés rencontrées pour le Forming des éléments mémoires OxRAM ne nous ont pas permis d'obtenir une forte répétabilité des résultats. Les observations ont dû être faites sur un nombre réduit de cellules en considérant les distributions cycle à cycle, et donc en ignorant la dispersion due à la fabrication de nombreuses cellules. De plus, l'effet mémoire peut avoir perturbé les observations obtenues lorsque l'on a fait varier des paramètres d'écriture au cours de la vie de la cellule. C'est pourquoi, dans le prochain chapitre, nous développerons puis exploiterons un circuit intégré spécifique pour approfondir cette question d'écriture en courant.

Bibliographie

[9] P. Jain *et al.*, « 13.2 A 3,6 Mb 10,1 Mb/mm2 Embedded Non-Volatile ReRAM Macro in 22 nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0,5 V with Sensing Time of 5 ns at 0,7 V », in *2 019 IEEE International Solid- State Circuits Conference - (ISSCC)*, févr. 2019, p. 212-214. doi: 10.1109/ISSCC.2019.8662393.

[12] B. Giraud *et al.*, « Benefits of Design Assist Techniques on Performances and Reliability of a RRAM Macro », in *2 023 IEEE International Memory Workshop (IMW)*, mai 2023, p. 1-4. doi: 10.1109/IMW56887.2023.10145984.

[16] G. Sassine, C. Cagli, J.-F. Nodin, G. Molas, et E. Nowak, « Novel Computing Method for Short Programming Time and Low Energy Consumption in HfO2 Based RRAM Arrays », *IEEE Journal of the Electron Devices Society*, vol. 6, p. 696-702, 2018, doi: 10.1109/JEDS.2018.2830999.

[25] W.-H. Chen *et al.*, « A 16 Mb dual-mode ReRAM macro with sub-14 ns computing-in-memory and memory functions enabled by self-write termination scheme », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 28. doi: 10.1109/IEDM.2017.8268468.

[32] A. Glukhov *et al.*, « Statistical model of program/verify algorithms in resistive-switching memories for in-memory neural network accelerators », in *2 022 IEEE International Reliability Physics Symposium (IRPS)*, mars 2022, p. 3. doi: 10.1109/IRPS48227.2022.9764497.

[48] F. M. Puglisi, N. Zagni, L. Larcher, et P. Pavan, « Random Telegraph Noise in Resistive Random Access Memories: Compact Modeling and Advanced Circuit Design », *IEEE Transactions on Electron Devices*, vol. 65, nº 7, p. 2964-2972, juill. 2018, doi: 10.1109/TED.2018.2833208.

[51] Z.-Q. Wang *et al.*, « Cycling-induced degradation of metal-oxide resistive switching memory (RRAM) », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 7. doi: 10.1109/IEDM.2015.7409649.

[54] A. Grossi *et al.*, « Fundamental variability limits of filament-based RRAM », in *2 016 IEEE International Electron Devices Meeting (IEDM)*, déc. 2016, p. 4. doi: 10.1109/IEDM.2016.7838348.

[66] B. Chen *et al.*, « A Novel Operation Scheme for Oxide-Based Resistive-Switching Memory Devices to Achieve Controlled Switching Behaviors », *IEEE Electron Device Letters*, vol. 32, nº 3, p. 282-284, mars 2011, doi: 10.1109/LED.2010.2101577.

[67] W. Lian *et al.*, « Improved Resistive Switching Uniformity in $ \hboxCu/HfO_2/\hboxPt$ Devices by Using Current Sweeping Mode », *IEEE Electron Device Letters*, vol. 32, nº 8, p. 1053-1055, août 2011, doi: 10.1109/LED.2011.2157990.

[80] A. Padovani, J. Woo, H. Hwang, et L. Larcher, « Understanding and Optimization of Pulsed SET Operation in HfOx-Based RRAM Devices for Neuromorphic Computing Applications », *IEEE Electron Device Letters*, vol. 39, nº 5, p. 672-675, mai 2018, doi: 10.1109/LED.2018.2821707.

[81] J. Reuben, M. Biglari, et D. Fey, « Incorporating Variability of Resistive RAM in Circuit Simulations Using the Stanford–PKU Model », *IEEE Transactions on Nanotechnology*, vol. 19, p. 508-518, 2020, doi: 10.1109/TNANO.2020.3004666.

[82] P. R. Shrestha *et al.*, « Analysis and Control of RRAM Overshoot Current », *IEEE Transactions on Electron Devices*, vol. 65, nº 1, p. 108-114, janv. 2018, doi: 10.1109/TED.2017.2776860.

[83] P. Trotti, « Dispositifs et circuits à convenance énergétique pour les mémoires résistives », 2022. [En ligne]. Disponible sur: http://www.theses.fr/2022GRALY073/document

[84] C. Cagli *et al.*, « Study of the Energy Consumption Optimization on RRAM Memory Array for SCM Applications », in *2017 IEEE International Memory Workshop (IMW)*, 2017, p. 1-4. doi: 10.1109/IMW.2017.7939106.

[85] G. Piccolboni *et al.*, « Investigation of Cycle-to-Cycle Variability in HfO2-Based OxRAM », *IEEE Electron Device Letters*, vol. 37, nº 6, p. 721-723, juin 2016, doi: 10.1109/LED.2016.2553370.

Chapitre IV : Écriture en courant sur une matrice mémoire

IV.1 Introduction

Après l'expérimentation de l'écriture en courant sur le circuit *WT-PAOLA* présenté dans le chapitre précédent, nous souhaitons dépasser les limitations auxquelles nous avons fait face. C'est l'objet de ce chapitre.

Comme le principal facteur limitant était la conception du circuit *WT-PAOLA*, dont nous avons fait un usage détourné, la première étape est de concevoir un circuit spécifique et dédié à notre expérience. Nous commençons donc par établir un ensemble d'objectifs que ce circuit devra remplir, puis nous discutons de l'architecture finale que l'on a sélectionnée pour ce circuit que l'on appellera *MEMORIIC* dans la suite. La conception du circuit *MEMORIIC* n'a pas été réalisée durant cette thèse, mais durant le stage de fin d'études l'ayant précédée, nous ne rentrerons donc pas dans les détails ici. Cependant, les informations utiles à la compréhension de la suite de ce chapitre seront répétées.

Nous discutons ensuite de la conception et de la réalisation de l'environnement d'exploitation du circuit *MEMORIIC*. Cela inclut à la fois une partie matérielle et une partie logicielle. Les différents choix, parfois peu communs, effectués lors de cette conception, ainsi que leur motivation et leur conséquence sont discutés. Les méthodes d'écriture implémentées et utilisées par la suite sont également présentées.

Enfin, nous nous intéresserons aux résultats concernant l'écriture en courant obtenus par l'exploitation du circuit *MEMORIIC*. Grâce à la conception du circuit *MEMORIIC*, nous pouvons présenter des résultats comparant directement sur la même puce différentes méthodes d'écriture en courant, mais également en tension sur un large ensemble de cellules 1T1R. Une première analyse du FORMING sera présentée, avant de considérer le cas nous intéressant le plus, celui du SET.

IV.2 Conception du circuit *MEMORIIC*

Le circuit *MEMORIIC* a été réalisé pendant le stage de fin d'études précédant cette thèse. Son dessin (« layout » en anglais) a quant à lui été réalisé pendant cette thèse. Des éléments du travail réalisé pendant le stage sont répétés ici pour plus de clarté, mais les éléments de détails ne le sont pas.

IV.2.1 Objectifs

Les expériences menées sur le circuit *WT-PAOLA* dans le Chapitre III ont mis en lumière le rôle très important joué par la capacité parasite C_p dans l'écriture en courant. Ainsi, la co-intégration de la source de courant aux éléments mémoires OxRAM est essentielle pour pouvoir obtenir des résultats représentatifs du cas d'application intégré de l'écriture en courant. L'option de la simulation a été

écartée par manque de modèle adapté, comme discuté dans l'introduction du Chapitre III. Ce besoin de données expérimentales est la principale motivation derrière la réalisation du circuit *MEMORIIC*.

Les résultats du chapitre précédent souffraient de quelques limitations que nous souhaitons régler à l'aide de ce nouveau circuit.

1. **Difficultés à effectuer le Forming**
   Sur le circuit *WT-PAOLA*, l'étape de Forming était difficile à réaliser. Elle demandait un travail important et n'était pas bien contrôlée. Après l'opération de formation d'une cellule et de quelques cycles de tests, elle était soit exploitée car le Forming semblait s'être bien passé, soit elle était éliminée pour cause d'un mauvais Forming. Cependant, cette prise de décision n'était pas objective et a donc pu introduire un biais important sur le type de cellule finalement étudiée.
   Le circuit *MEMORIIC* doit donc permettre d'éviter cela. Pour cela, on souhaiterait pouvoir effectuer un Forming standard, c'est-à-dire avec une source de tension. Il est donc nécessaire d'avoir une **limitation de courant** intégrée, ainsi que la possibilité d'effectuer une **écriture en tension**.
2. **Statistique limitée**
   De part à la fois le Forming difficile, mais également de part la conception du circuit *WT-PAOLA*, le nombre de cellules exploitables était faible. En effet, chaque circuit *WT-PAOLA* ne comporte que 4 cellules 1R. Ainsi, dans les résultats obtenus dans le précédent chapitre, nous nous sommes concentrés sur les statistiques cycle-à-cycle d'une cellule. Cela présente trois inconvénients. Le premier est l'impossibilité d'étudier la statistique cellule-à-cellule qui est très importante dans la réalisation d'une mémoire pouvant contenir des millions de cellules. Le second est que les résultats comparatifs présentés ont pu être influencés par l'ordre dans lequel les différents scénarios ont été réalisés. En effet, l'effet mémoire peut avoir biaisé les résultats. Enfin, cela permet de vérifier si l'effet de la capacité d'interconnexion C_p engendre des non-uniformités corrélées à la localisation spatiale des cellules.
   Pour le circuit *MEMORIIC*, nous souhaitons donc avoir un large ensemble de cellules sur lequel effectuer nos expériences, cela requiert donc l'utilisation d'une **matrice mémoire** permettant de multiplexer les lignes d'accès aux éléments mémoires OxRAM par un système d'adressage. L'implémentation d'un tel multiplexage est facilitée par l'utilisation de **cellules 1T1R**. Le transistor de la cellule 1T1R pourra également être utilisé pour limiter le courant.
3. **Une unique méthode d'écriture**
   Le circuit *WT-PAOLA* n'offrait pas la possibilité d'écrire les éléments mémoires OxRAM avec une source de tension. Ainsi, nous n'avons pas pu comparer notre écriture en courant avec d'autres méthodes d'écriture. La comparaison que nous avons pu effectuer se repose sur des données extraites d'un autre circuit, possiblement dans des conditions assez différentes des nôtres, mais surtout hors de notre contrôle.
   Pour pouvoir juger de l'intérêt d'une technique d'écriture, il est plus aisé de la comparer à une autre dans une situation la plus identique possible. Dans l'idéal, la comparaison serait effectuée entre deux ensembles d'élément mémoire OxRAM au sein d'une même matrice (sur le même circuit par extension) et avec le même historique d'opérations. La différence due à la variation statistique entre les cellules peut alors être compensée par l'utilisation d'un grand nombre de cellules dans chaque ensemble.
   Pouvoir réaliser de telles comparaisons est un des objectifs du circuit *MEMORIIC*. Pour ce faire, le circuit *MEMORIIC* doit, en plus d'avoir une **matrice mémoire** intégrant un grand nombre de dispositifs, permettre aussi bien d'écrire les éléments mémoires OxRAM à l'aide d'une **source de courant** que d'une **source de tension**.
4. **Environnement expérimental contraignant**
   La mise en place du banc de mesure, ainsi que les mesures du circuit *WT-PAOLA* ont été effectuées sous pointes. Le test sous pointes pour des circuits complets n'est pas aisé. Premièrement, le

> nombre de connexions électriques externes est limité, ce qui limite la conception du circuit et nécessite souvent différents modes de multiplexage. Deuxièmement, la pose des pointes use les plots métalliques, cela a pour conséquence de limiter le nombre d'utilisations d'un même circuit. La possibilité d'avoir une connexion mal établie entre une des pointes et son plot augmente le nombre d'incertitudes dès qu'une bizarrerie est observée. Et troisièmement, le temps de mesure disponible est faible car l'équipement est onéreux. Le partage des équipements requiert également l'installation et la désinstallation de l'ensemble du dispositif de test à chaque nouveau créneau. On passe donc beaucoup de temps à faire et défaire, vérifier que l'on ne s'est pas trompé et que l'on a bien reproduit la même mise en place, rechercher l'instrument qui a été déplacé pour être utilisé ailleurs, changer d'équipement car celui précédemment utilisé est indisponible, reconfigurer les différents instruments ou courir le risque d'avoir une mauvaise surprise, etc.
> Par conséquent, nous souhaitons que le circuit *MEMORIIC* soit mis en boîtier, et, dans la limite du raisonnable, que son exploitation puisse être réalisée à l'aide d'un circuit imprimé sans nécessiter d'appareil externe.

Au-delà des limitations du circuit *WT-PAOLA*, nous souhaitons également que le circuit *MEMORIIC* conserve certaines de ses fonctionnalités. C'est particulièrement le cas de la lecture. Dans les implémentations intégrées, la lecture est souvent effectuée par comparaison du courant traversant une cellule avec celui traversant une référence. L'information obtenue est alors binaire. Il est possible de faire varier cette référence pour obtenir, par lecture successive, une information plus résolue de l'état de conductance de l'élément mémoire OxRAM. Sur le circuit *WT-PAOLA*, la lecture est effectuée par mesure du courant traversant un élément mémoire OxRAM sous une tension fixée à 0,2 V à l'aide d'un instrument de laboratoire externe. Cette **mesure analogique** nous a permis d'obtenir une quantité d'informations plus importante sur la conductance. Cela rend les données plus simples à interpréter et permet d'observer des phénomènes tels que la relaxation qui sont difficilement observables avec des systèmes de lectures successives espacées dans le temps. Pour le circuit *MEMORIIC*, nous souhaitons conserver cette capacité d'obtenir une lecture analogique. Dans le cas du circuit *WT-PAOLA*, ce type de mesure était la conséquence naturelle de l'environnement expérimental type laboratoire utilisé. Dans le cas du circuit *MEMORIIC* ainsi que son environnement expérimental plus limité, un soin particulier devra être apporté pour permettre une lecture analogique.

Ainsi les objectifs du circuit *MEMORIIC* peuvent être résumés en :

1. Intégrer une matrice de cellules 1T1R pour permettre l'obtention de résultats statistiquement représentatifs
2. Y co-intégrer une source de courant programmable couvrant une large plage de courant permettant leur écriture
3. Permettre l'écriture des éléments mémoires OxRAM par une source de tension, telles que le font les méthodes de l'état de l'art
4. Permettre une lecture analogique de la conductance de chaque élément mémoire OxRAM
5. Prévoir une mise en boîtier et un circuit imprimé compagnon le plus autonome possible

IV.2.2 Architecture générale

Concernant le premier objectif, nous réutilisons une matrice de 16k (2^{14}) cellules 1T1R précédemment réalisée au CEA-Leti. Elle permet l'accès aux éléments de façon unitaire avec un adressage sur 14 bits (A_{0-13}). Le transistor de la cellule 1T1R sert de transistor de sélection pour l'adressage, mais il peut également être utilisé comme limiteur de courant lorsque cela est nécessaire.

Les éléments mémoires OxRAM ont un fonctionnement bipolaire, ainsi, la source doit être alternativement connectée à l'une des électrodes de conduction de la cellule, puis à l'autre pour pouvoir effectuer le SET et le RESET. Pour simplifier à la fois la description de l'architecture, mais également son implémentation, nous utilisons un circuit permettant d'intervertir la connexion entre deux entrées et

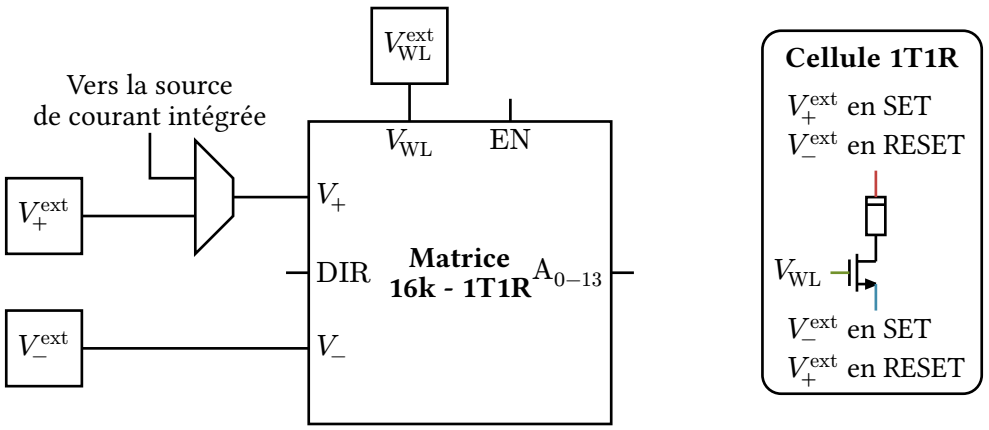


Figure 4.1 – Interface de la matrice mémoire 16k et cellule 1T1R.

deux sorties par commande numérique. Ainsi, comme illustré sur la Figure 4.1, les entrées analogiques de la matrice correspondant aux connections faites avec la cellule sélectionnée sont la tension haute V_+ connectée au terminal libre de l'élément mémoire OxRAM en SET, et à la source du transistor de la cellule en RESET, la tension basse V_- connectée au terminal opposé à celui connecté à V_+, et la tension de grille V_{WL} connectée à la grille du transistor de la cellule sélectionnée.

Pour remplir le troisième objectif, qui consiste à être en capacité d'écrire les éléments mémoires OxRAM à l'aide d'une source de tension, nous prévoyons un mode de fonctionnement connectant V_+ à un plot extérieur V_+^{ext}. Nous avons fait le choix de ne pas intégrer la source de tension. En effet, contrairement à la source de courant, la source de tension n'est que peu affectée par la capacité parasite C_p. Ainsi elle proviendra de dispositifs externes au circuit qui seront intégrés au circuit imprimé décrit plus tard dans la partie IV.3.1. Ainsi, V_+ peut être soit connecté à la source de courant intégrée, soit à V_+^{ext} à l'aide d'un multiplexeur analogique. V_- et V_{WL} sont également connectés à l'extérieur du circuit intégré par l'intermédiaire des plots extérieurs V_-^{ext} et V_{WL}^{ext} (voir Figure 4.1).

Cette possibilité de connecter V_+ à l'extérieur du circuit au travers du plot V_+^{ext} nous permet également de réaliser la lecture analogique, demandée par le quatrième objectif, de façon externe. Dans le cas de la lecture, bien que la capacité parasite C_p pourrait ralentir la lecture en rallongeant la durée du régime transitoire, nous préférons réaliser la lecture de façon externe. Ne pas intégrer le système de lecture simplifie le circuit *MEMORIIC* et permet donc de réduire le risque de dysfonctionnement. Dans notre cas, nous estimons que ce gain de robustesse vaut mieux que le possible gain de vitesse d'une solution de lecture intégrée. En effet, nos ressources pour développer ce circuit sont limitées, et la lecture n'est pas l'objet du circuit *MEMORIIC*, ce n'est qu'un outil permettant de réaliser les expériences qui nous intéressent. Une solution de lecture intégrée au circuit imprimé sera développée pour respecter notre cinquième objectif.

De façon générale, nous essayons de minimiser le nombre de fonctions intégrées. Pour ce faire, une importante partie des signaux est connectée ou connectable à l'extérieur du circuit pour permettre une implémentation externe de la fonction, comme V_+, V_- et V_{WL} par exemple. Dans le cas de la référence de courant, une implémentation semi-intégrée (utilisant une résistance externe) a été réalisée, toutefois elle est également contournable. De même, la partie numérique intégrée est minimale, l'interface numérique du circuit est faite pour être directement commandée par un micro-contrôleur ou un FPGA. Aucun signal numérique de l'interface n'est sensible au timing ou aux fronts comme ce serait le cas avec une communication série ou avec un compteur. Cela nous évitera le problème rencontré sur le circuit *WT-PAOLA* avec le compteur et l'horloge. Une partie des bits numériques est multiplexée à l'interface. Cela permet de réduire le nombre de connexions externes. Néanmoins, elles le sont à l'aide de simple

bascule à deux états, l'un laissant le signal d'entrée passer de façon asynchrone, l'autre mémorisant et fixant leur valeur de sortie.

D'autres fonctionnalités (souvent de débogage) sont intégrées dans le circuit. Comme elles n'ont pas d'intérêt pour la suite de ce manuscrit, elle ne sont pas décrites ici.

IV.2.3 Source de courant

Le dernier élément intégré d'intérêt pour la suite est la source de courant programmable. Le second objectif requiert qu'elle couvre une large gamme de courant. Nous visons initialement une couverture commençant sous 1 µA et allant jusqu'à 1 mA. Cette dynamique a été choisie par rapport à l'expérience précédemment acquise concernant les régimes d'utilisation possible des éléments mémoires OxRAM avec lesquels nous travaillons. Pour rappel, les courants utilisés dans le circuit *WT-PAOLA* était compris entre 7,5 µA et 63 µA. Cette couverture incorpore une marge importante par rapport à la dynamique dont nous pensons avoir réellement besoin. Cette marge nous permet d'être certain que la source ne soit pas un élément limitant, même si les éléments mémoires OxRAM reçoivent des modifications changeant leur gamme de fonctionnement avant leur fabrication, ou qu'un autre problème vienne la limiter.

Pour une source de courant variable, couvrir une large dynamique signifie également avoir une grande précision et résolution par rapport au courant maximum de sortie. En effet, le courant minimum doit au moins représenter une unité de la résolution choisie. La résolution devrait donc dans notre cas être d'au moins 8 bits et la précision suffisante pour distinguer les plus faibles courants entre eux. Mais nous n'avons pas besoin d'avoir autant de précision ni autant de résolution autour de 1 µA qu'autour de 500 µA par exemple. Ainsi, pour relaxer les contraintes, nous avons décidé de découper la source de courant en plusieurs sources de courant intermédiaires chacune dimensionnée pour un intervalle différent. Pour couvrir efficacement toute la plage souhaitée, nous répartissons les intervalles de manière logarithmique.

Dans le cas des algorithmes de programmation utilisant l'adaptation et la vérification, il est intéressant d'avoir une source monotone par rapport à sa commande. Cela permet de s'assurer que lorsque l'algorithme pense augmenter l'amplitude, il augmente réellement. Pour assurer cette monotonie, nous nous orientons sur une architecture thermométrique. Dans ces architectures, la source de courant programmable est divisée en N sources de courant élémentaires fixes, généralement de même valeur. La commande correspond alors au nombre de sources élémentaires connectées à la sortie. L'ordre dans lequel elles sont connectées est prédéfini et stable. C'est ce principe qui assure la monotonie malgré les non-idéalités qui génèrent une erreur aléatoire sur le courant de chaque source élémentaire. Dans notre cas, il faut aussi assurer la monotonie lors d'un changement d'intervalle. Pour ce faire, nous appliquons le principe de l'architecture thermométrique aux deux échelles : au sein de chaque source de courant intermédiaire, ainsi qu'entre elles.

Du fait de l'intégration de la matrice 1T1R de 16k cellules, nous avons déjà la nécessité d'avoir un bus de 14 bits (A_{0-13}) dans l'interface numérique. Pour limiter le nombre de connexions externes nécessaires, nous souhaitons multiplexer ce bus avec la commande de la source de courant. Nous partons donc du principe que la commande de la source de courant s'effectue sur 14 bits ou moins (I_{0-13}).

Pour des raisons d'exploitation pratique du circuit, nous souhaitons que le courant généré soit proportionnel à la valeur de commande I_{0-13}. Pour simplifier le décodage, nous répartissons les intervalles couverts par les sources intermédiaires, de telle sorte qu'à chaque doublement de la valeur de commande l'intervalle change. Nous décidons de découper chaque intervalle en 16 valeurs soit 4 bits. Par conséquent, chaque source de courant intermédiaire est directement commandée par 4 bits de I_{0-13}, sans besoin de décodage supplémentaire. Le bit de poids juste supérieur à ces 4 bits est utilisé comme signal d'activation. Ce signal d'activation est également utilisé comme signal de forçage pour saturer

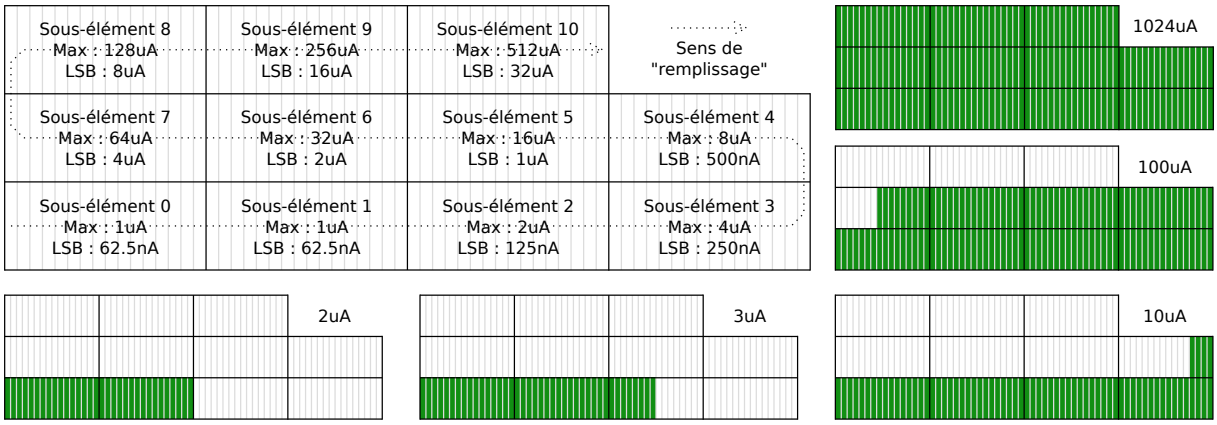


Figure 4.2 – Représentation de l'architecture de la source de courant intégrée et de l'ordre d'utilisation des sources unitaires pour différents exemples de courant de sortie.

la sortie des sources intermédiaires inférieures lorsqu'il est actif. Cet arrangement permet de réaliser l'architecture thermométrique à deux échelles avec peu de décodage et une architecture régulière. La seule exception est la plus petite source intermédiaire qui a son signal d'activation toujours actif. Avec cette architecture, nous pouvons commander 11 sources de courant intermédiaires avec notre bus de 14 bits. Comme $2^{10} = 1024$, 10 sources intermédiaires suffiraient pour couvrir la plage 1 µA à 100 µA initialement souhaitée. La 11e source intermédiaire disponible est utilisée comme plus faible source et permet de faire le pont entre 0 µA et 1 µA. La Figure 4.2 illustre le fonctionnement de cette architecture et précise les gammes de courant et le pas de courant au sein de chaque source intermédiaire.

Chaque source de courant intermédiaire est réalisée à l'aide d'un miroir de courant à 16 branches de sortie. Leurs courants de références sont dérivés d'un courant de référence global au travers d'un autre miroir de courant qui sert aussi de gain (en connectant plusieurs branches du miroir en parallèle). L'effet de cette architecture de dérivation des courants sur l'erreur due aux variations aléatoires de la caractéristique des transistors (« mismatch » en anglais) est détaillé dans le rapport de stage. Les différents miroirs ont été dimensionnés en conséquence.

IV.3 Conception de la plateforme de test

IV.3.1 Partie matérielle

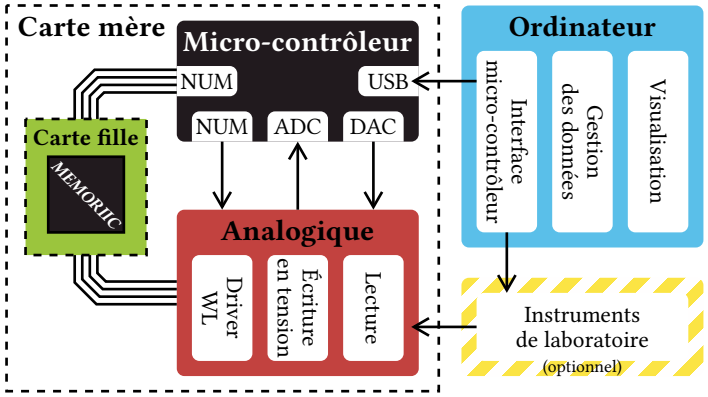


Figure 4.3 – Schéma de principe de la plateforme de test du circuit *MEMORIIC*.

Nous avons choisi de mettre en boîtier notre circuit intégré. Ce choix nous a permis d'avoir un plus grand nombre de connexions externes que nous avons mis à profit pour simplifier notre circuit ainsi que pour assurer des solutions de repli en cas de dysfonctionnement d'une partie non essentielle du circuit. La mise en boîtier nous permet également de ne pas nécessiter l'utilisation de plateforme de test sous pointe, nous libérant d'un problème logistique important. Pour aller jusqu'au bout de la

démarche, nous souhaitons, dans la limite du raisonnable, nous passer des instruments de laboratoire dont la disponibilité n'est pas assurée. Après tout, de tels instruments sont faits pour la métrologie, toutefois, nous n'avons pas besoin d'avoir une grande précision sur nos observations qui sont déjà fortement perturbées par la variabilité des éléments mémoires OxRAM, ainsi que par les bruits qu'elles génèrent[29]. De plus nous avons pour objectif d'effectuer des observations comparatives dont la valeur absolue n'a pas de grande importance. C'est pourquoi nous avions établi comme cinquième objectif d'avoir un circuit imprimé compagnon le plus indépendant possible.

La Figure 4.3 illustre la topologie générale de la plateforme de test développée pour l'exploitation du circuit *MEMORIIC*.

Le circuit *MEMORIIC* est mis en boîtier et soudé sur une carte fille qui permet de réutiliser la même carte mère pour toutes les puces, sans avoir besoin de les dé-souder. La carte fille ne comporte que les capacités de découplage et peut donc être produite en nombre suffisant. La solution alternative consistant à utiliser des douilles (« sockets » en anglais) n'a pas été retenue ici par crainte du manque de fiabilité, particulièrement après un grand nombre de changements de puce et un retour d'expérience négatif. Les douilles requièrent une mise en place soigneuse laissant planer le doute sur la présence ou non de faux contacts lors du débogage d'un problème. Elles ont l'avantage d'offrir une meilleure intégrité du signal (en l'absence de faux contact). Cependant, en l'absence de signaux sensibles aux timing ou aux fronts ou de signaux radio-fréquence, l'intégrité des signaux n'est pas un critère important pour nous, et l'utilisation d'une carte fille est suffisante pour répondre à nos besoins.

La carte mère comporte quant à elle l'ensemble des circuits externes au circuit intégré. Cela inclut les circuits analogiques externes nécessaires à l'écriture en tension des éléments mémoires OxRAM, à leur lecture et à la polarisation de la grille des sélecteurs (V_{WL}). L'interface numérique est réalisée à l'aide d'un microcontrôleur. Certaines de ses sorties numériques ainsi que ses convertisseurs analogiques vers numérique (ADC) et numérique vers analogique (DAC) sont également utilisés pour interfacer avec la partie analogique de la carte mère. Le microcontrôleur communique par USB avec un ordinateur au travers d'une émulation d'une communication série. L'ordinateur est responsable de la coordination, il donne les ordres au microcontrôleur, récupère, stocke et organise les données mesurées et il permet leur visualisation.

Cette architecture permet une exploitation directement « au bureau » car toutes les fonctionnalités sont intégrées soit dans le circuit *MEMORIIC*, soit sur la carte mère, soit dans l'ordinateur. Cependant, il est également possible de l'associer à des instruments de laboratoire. Par exemple, la calibration des ADC/DAC a été réalisée avec un multimètre de laboratoire. La carte mère comporte un certain nombre de cavaliers permettant cette flexibilité. La quasi-totalité des signaux analogiques du circuit *MEMORIIC* peut ainsi être couplée à des instruments de laboratoire par le biais de connecteurs SMA ou être connectée aux fonctions internes de la carte mère pour un fonctionnement autonome.

Concernant le circuit de lecture, la Figure 4.4 montre trois étapes du choix de l'architecture utilisée. La première est le schéma de principe de ce que le circuit de lecture doit accomplir. Il doit permettre d'imposer une tension fixe (dans notre cas 0,2 V) sur la cellule intégrée et de mesurer le courant la traversant. Généralement, pour réaliser une telle mesure, une résistance « shunt » R_{mes} est utilisée, cela correspond au deuxième schéma. Pour identifier le problème apporté par cette solution, essayons de dimensionner R_{mes}. D'un côté, R_{mes} doit être très inférieur à la résistance équivalente de la cellule mesurée pour que la presque totalité de la tension soit bien sur la cellule. La résistance d'un élément mémoire OxRAM peut être aussi faible que 5 kΩ. Il faudrait donc avoir R_{mes} inférieur à moins de 50 Ω

[29] Si l'étude de ses bruits, ou de la variabilité des éléments mémoires OxRAM était au cœur de nos préoccupassions, une attention plus importante devrait être apportée à la méthode et aux instruments de mesure utilisés. Cependant, nous ne cherchons qu'à comparer des méthodes de programmation.

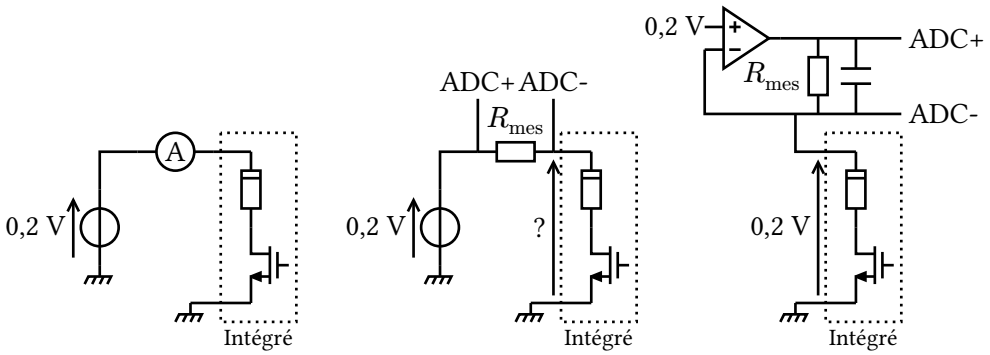


Figure 4.4 – De gauche à droite, schéma de principe d'un circuit de lecture, schéma d'une méthode de mesure de courant typique et schéma du circuit de lecture utilisé et intégré sur la carte mère.

pour avoir moins de 1 % d'incertitude sur la tension aux bornes de la cellule. Cela donnerait alors un signal de 50 µV pour un courant de lecture de 1 µA. On pourrait utiliser un amplificateur pour amplifier ce signal avant de le numériser, mais nous serions confrontés à l'offset de l'amplificateur. Nous avons donc opté pour la solution illustrée par le troisième schéma. A l'aide d'une boucle de rétroaction, la tension de lecture est régulée à 0,2 V tout en conservant R_{mes} sur le chemin du courant. La rétroaction a pour effet de présenter une faible impédance du point de vue de la cellule mesurée. Ainsi la tension de mesure est bien définie et la tension générée par R_{mes} est adaptée aux ADC car elle peut être choisie grande (30 kΩ dans notre cas). Une contrainte de ce montage est le risque d'instabilité. Une capacité en parallèle de R_{mes} est utilisée pour assurer la stabilité.

La carte mère ainsi que la carte fille ont été réalisées avec l'aide de Vincent Josselin ingénieur-chercheur au CEA-Leti. Il a en particulier réalisé la conception de la partie numérique de la carte mère ainsi que l'alimentation. Il a également produit les documents CAD, réalisé le dessin des masques, déterminé et acheté les composants et organisé la sous-traitance pour la réalisation des cartes ainsi que leur montage.

IV.3.2 Partie logicielle

La partie logicielle couvre à la fois le programme embarqué par le microcontrôleur, un STM32G474, et celui présent sur l'ordinateur.

La première motivation derrière la plateforme logicielle développée est avant tout la volonté d'avoir un historique intégral des opérations effectuées sur chaque cellule. Cet historique devrait contenir aussi bien les éventuels paramètres des opérations effectuées que les résultats auxquels elles auraient conduit. L'intérêt est double. D'un côté, cela nous permet de connaître précisément l'histoire des cellules que l'on utilise, et donc d'éviter de subir un biais dû à une expérience passée et oubliée qui les aurait affectées. D'un autre côté, cela nous permet de pouvoir connaître avec précision les conditions d'obtention des différents résultats, facilitant ainsi leur analyse, même *a posteriori*.

Un historique n'est intéressant que s'il est précis, ainsi, il ne doit pas être maintenu manuellement. Sa gestion doit être intégrée au logiciel orchestrant les expériences. Ce choix est donc le point de départ de notre développement. De plus, nous souhaitons que l'utilisateur contrôle la plateforme à l'aide d'une interface graphique permettant la communication dans les deux sens entre l'utilisateur et l'ordinateur.

L'utilisation d'un micro-contrôleur en combinaison avec un ordinateur pose la question de la répartition des responsabilités. Une partie de ce que l'ordinateur fait pourrait être fait par le micro-contrôleur. Bien sûr un certain nombre de différences entre les deux en fixe certaines limites. Par exemple,

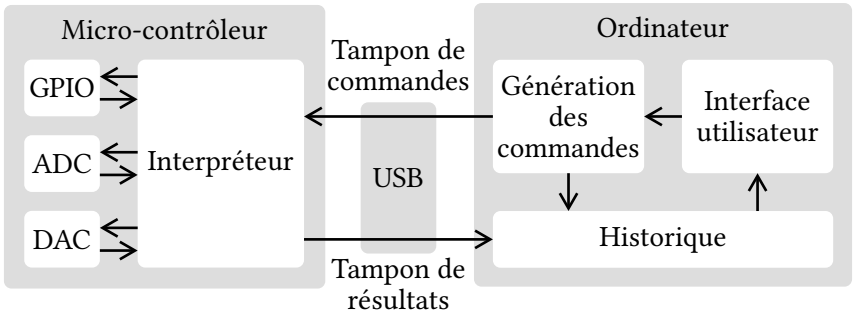


Figure 4.5 – Représentation des différents composants logiques, de leur attribution à l'ordinateur ou au micro-contrôleur et de leur communication. L'ordinateur et le micro-contrôleur fonctionnent de façon asynchrone.

l'ordinateur est le seul capable d'assurer le stockage des données nécessaires à notre historique car le stockage disponible sur le micro-contrôleur n'a pas assez de capacité. D'un autre côté, le micro-contrôleur est le seul à pouvoir commander ses périphériques en respectant un timing précis. Nous concluons qu'il est donc nécessaire que l'ordinateur soit informé de toutes les opérations élémentaires effectuées, pour qu'il puisse établir l'historique, et que le micro-contrôleur soit capable d'enchaîner les opérations en autonomie pour pouvoir respecter un timing. Comme l'utilisateur communique avec l'ordinateur, c'est l'ordinateur qui doit être à l'initiative des opérations exécutées.

La solution retenue pour minimiser les aller-retours de données et respecter cette répartition minimale des rôles est illustrée sur la Figure 4.5. Un ensemble de micro-commandes constituant un programme sont générées sur l'ordinateur et peuvent ainsi être directement intégrées à l'historique. Elles sont transmises au micro-contrôleur dans un tampon contenant l'intégralité du programme. Le micro-contrôleur exécute alors ce programme en réalisant chaque micro-commande les unes après les autres, en autonomie, sans aucune synchronisation avec l'ordinateur. Cela lui permet de pouvoir respecter le timing entre les micro-commandes. Il stocke temporairement les résultats dans sa mémoire et initie une transmission en bloc à l'ordinateur dès que sa capacité de stockage est proche de sa limite. Après avoir fini l'exécution du programme, le micro-contrôleur le communique à l'ordinateur et attend la transmission d'un nouveau programme. C'est à ce moment là que le micro-contrôleur et l'ordinateur se re-synchronisent. La responsabilité du micro-contrôleur est donc essentiellement réduite à celle d'un interpréteur de commandes. Toutes les autres sont attribuées à l'ordinateur.

Cette répartition déplace une large partie de l'intelligence et donc du code, du côté de l'ordinateur. Cela est idéale puisqu'il est plus facilement programmable, offre plus de puissance de calcul et de mémoire, et les outils tels que les débugueurs sont plus accessibles. Cela permet également de proposer à l'utilisateur plusieurs niveaux de commandes, d'un contrôle des plus directs au travers de programmes constitués d'une seule commande élémentaire, par exemple, modifier la valeur logique d'un bit, au plus complexe incluant des millions de commandes, comme réaliser une multitude de cycles d'écriture vérifiés sur l'ensemble des cellules de la matrice.

Comparée à l'utilisation de macro-commandes[30] remplaçant un programme entier en une seule commande, notre solution utilisant des micro-commandes requiert un plus grand investissement dans le développement des commandes. En effet, la taille des programmes est limitée par la capacité de

[30]Ici, on entend par macro-commandes des commandes représentant l'équivalent d'un programme complet dont le détail serait défini par le logiciel du micro-contrôleur, en opposition avec les micro-commandes que nous utilisons qui représentent une action élémentaire. Des exemples de macro-commandes seraient l'écriture, l'effacement, la lecture, voire la réalisation d'une multitude de cycles sur une cellule ou sur un groupe de cellules. Au contraire, des exemples de micro-commandes sont la modification d'un bit d'un GPIO, l'attente d'un certain temps, la modification de la valeur d'un DAC ou encore le démarrage d'une lecture d'un ADC.

la mémoire du micro-contrôleur (qui doit être répartie entre les tampons pour la communication USB, le tampon de commandes, les registres de l'interpréteur et les tampons de résultats). Comme le micro-contrôleur se re-synchronise avec l'ordinateur à chaque fin de programme, il n'est pas possible de simplement découper un programme en deux car le timing n'est alors plus respecté. L'ordre de grandeur d'une opération d'écriture est de 10 µs alors que l'ordre de grandeur d'une synchronisation entre le micro-contrôleur et l'ordinateur est de 10 ms. Lorsqu'une opération doit être exécutée sur 2k cellules, la durée totale est de l'ordre de 20 ms si aucune synchronisation est nécessaire et de 20 s sinon. L'exécution est fortement ralentie par cette synchronisation, ce qui augmente le temps de mesure inutilement. Pour réaliser l'équivalent d'un plus grand nombre de commandes dans le même espace de stockage[31], il est alors nécessaire d'inclure des commandes de saut pour permettre la construction de boucles et inclure également une mémoire pour l'interpréteur sous forme de registre ainsi que des commandes pour les manipuler pour que chaque itération de boucle puisse produire des résultats différents. Le détail des micro-commandes développées est donné en Annexe A.

Concernant la partie logicielle de l'ordinateur, le choix a été fait de la construire de manière monolithique, en opposition à modulaire. L'intérêt d'une telle construction est de ne pas avoir eu à faire des choix de découpage en différents modules tôt dans le développement. En effet, les caractéristiques et fonctionnalités précises souhaitées n'étaient pas connues à l'avance. Nous avons développé cette partie logicielle au fil de l'expérimentation, pour répondre à nos besoins au rythme où nous les découvrions. L'approche monolithique nous a permis de ne pas avoir à réconcilier des modules ayant divergé, car initialement distincts, après avoir découvert qu'ils devaient avoir une interaction plus resserrée. Au final une certaine modularité est apparue, cependant elle est plutôt la conséquence que la cause. Cette approche, peut-être peu scolaire, a également facilité la réalisation d'une interface utilisateur riche et dynamique mêlant commande de la plateforme, visibilité sur sont état en temps réel, visualisation des résultats tout juste acquis, en cours d'acquisition ou enregistrés dans l'historique. La Figure 4.6 montre une capture d'écran de l'interface du logiciel développé.



Figure 4.6 – Capture d'écran de l'interface du logiciel développé. On peut voir à gauche l'historique des opérations et les contrôles permettant de paramétrer et lancer une nouvelle opération, sur fond blanc une représentation graphique des opérations sélectionnées dans l'historique, en bas à droite une fenêtre permettant de voir l'intégralité de l'historique d'une cellule.

[31] Il devrait également être possible d'utiliser un flux pour communiquer les commandes au micro-contrôleur au rythme où il les consomme. Cependant, une telle solution nous paraît plus complexe à réaliser, elle n'a donc pas été envisagée.

0	1	2	3	4	5	6	7
1	0	3	2	5	4	7	6
7	6	5	4	3	2	1	0
6	7	4	5	2	3	0	1
2	3	0	1	6	7	4	5
3	2	1	0	7	6	5	4
5	4	7	6	1	0	3	2
4	5	6	7	0	1	2	3

Figure 4.7 – Répartition des huit groupes de cellules numérotées de 0 à 7. Chaque case mesure 16 cellules de côté et contient donc 128 cellules.

IV.3.3 Opération sur les cellules et algorithmes d'écriture

Dans cette section, nous décrivons les opérations effectuées sur les cellules du point de vue électrique à l'aide de la plateforme décrite précédemment. Nous nous concentrerons sur les techniques utilisées pour obtenir les données présentées dans la section suivante (Chapitre IV.4).

Pour permettre une meilleure comparaison entre les différentes méthodes, nous divisons la matrice de 16k cellules en groupes d'éléments mémoires OxRAM. Cela permet d'effectuer un test différent sur chaque groupe au sein d'une même puce, ce qui permet de comparer les résultats en éliminant les éventuels effets dus aux dérives dues au procédé de fabrication entre les différentes puce. Dans les résultats présentés, la matrice est divisée en huit groupes de 2k cellules, un nombre suffisantes pour obtenir une statistique cellule-à-cellule à 3σ. Les groupes sont réparties de façon à assurer une même distribution de leur membres sur chaque colonne et ligne de la matrice, ce qui évite d'introduire un éventuel biais dû à l'architecture de la matrice. La Figure 4.7 montre comment ces groupes sont réparties sur la matrice.

Les algorithmes présentés sont faits pour être appliqués sur des groupes entiers à la fois. Si l'on considère ces algorithmes comme une suite d'étapes, on peut organiser les choses de deux manières. Soit l'on effectue l'ensemble des étapes sur un élément mémoire OxRAM avant de passer au suivant, soit l'on effectue chaque étape sur l'ensemble des éléments mémoires OxRAM avant de passer à la suivante. Comme le changement d'adresse est plus rapide que le changement de la valeur des DAC, qui servent à définir V_{+}^{ext} et $V_{\text{WL}}^{\text{ext}}$, il semble plus intéressant d'effectuer chaque étape sur l'ensemble des éléments mémoires OxRAM avant de passer à la suivante. Ainsi, pour permettre une exécution plus rapide, les algorithmes sont toujours organisés de cette façon.

Tous les programmes sont constitués d'une suite d'opération d'écriture en courant ou en tension ou de lecture. Ces trois types d'opérations de base commencent par une phase de mise en place où le mode de fonctionnement du circuit *MEMORIIC* est correctement sélectionné pour correspondre à l'opération entreprise, et où les tensions des DAC contrôlant V_{+}^{ext} et $V_{\text{WL}}^{\text{ext}}$ sont fixées à la valeur demandée. L'adresse de la cellule à affecter est ensuite sélectionnée. La suite des opérations dépend de leur type.

1. **Lecture**

   Dans ce cas, $V_{\text{WL}}^{\text{ext}}$ est fixé à son maximum, 4,6 V. Le signal EN de la matrice est activé, ce qui permet au courant de circuler dans l'élément mémoire OxRAM sélectionné. Après un délai d'une 1 µs, l'acquisition de l'ADC du circuit de lecture est démarrée. Lorsqu'elle est terminée, le signal EN est désactivé.

2. **Écriture en tension**

   Le signal EN de la matrice est activé pendant le temps indiqué par l'utilisateur (entre 1 µs et 1 ms)

puis désactivé. Cela a pour effet d'exposer l'élément mémoire OxRAM à une impulsion de tension de la durée choisie.

3. **Écriture en courant**
   V_+ est connecté à la masse pour décharger la ligne électrique et s'assurer de commencer à 0 V. V_+ est ensuite déconnecté de la masse, puis le signal EN de la matrice et de la source de courant est activé pendant le temps indiqué par l'utilisateur. Ils sont ensuite désactivés.

L'adresse de la cellule sélectionnée est ensuite modifiée pour adresser la cellule suivante, et la partie de l'opération décrite juste au-dessus se répète (sans refaire la mise en place). L'opération prend fin après avoir affecté toutes les cellules du groupe visé.

Un système de masque constitué d'un bit par adresse permet de réaliser des écritures avec vérifications. Les opérations d'écriture peuvent être inhibées par ce masque, elles n'effectuent alors aucune action et passent directement à l'adresse suivante. Le masque est modifié par l'opération de lecture dès qu'une valeur de conductance dépassant un seuil est observée.

A partir de ces opérations de base, un ensemble de programmes paramétrables est réalisé. Ces programmes sont résumés dans le Tableau 4.1.

IV.4 Résultats expérimentaux

Dans cette section, nous présentons les résultats expérimentaux obtenus sur le circuit *MEMORIIC* à l'aide de la plateforme précédemment décrite. Contrairement au Chapitre III, toutes les distributions présentées dans ce chapitre sont cellule-à-cellule et non plus cycle-à-cycle.

IV.4.1 Formation du filament

La formation du filament conducteur est la première étape de la vie d'un OxRAM. Ainsi elle a le potentiel d'influencer ses performances pour toute la vie de la mémoire. Cette idée est supportée par l'état de l'art avec par exemple une démonstration d'une amélioration des performances en effectuant la formation à température élevée [86]. Cependant, d'autres publications [73] rapportent que les paramètres de leur algorithme de formation n'ont pas d'influence sur les performances de leurs dispositifs après cyclage. Nous allons donc essayer de **juger par nous-même de l'importance du Forming**. Dans le même temps, cela nous permet aussi de déterminer une condition de formation du filament pour nos prochaines expériences.

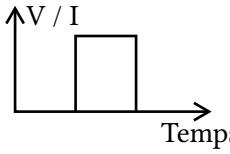
Nous en profitons pour essayer la formation du filament avec une source de courant. Comme indiqué dans le Chapitre III, lorsqu'une source de courant est utilisée, la présence de la capacité parasite C_p combinée à l'augmentation soudaine de la conductance de l'élément mémoire OxRAM produit un pic de courant. Dans le cas de la formation du filament, ce pic est attendu comme particulièrement fort. En effet, son amplitude maximum est proportionnelle à la variation de conductance. Or durant la formation, l'élément mémoire OxRAM passe d'un état essentiellement isolant à un état conducteur pouvant atteindre les 100 µS. Un pic d'environ 100 µA peut donc être attendu (en supposant la formation du filament à 1 V). Pour cette raison, nous décidons d'explorer trois principaux modes de formation des filaments avec :

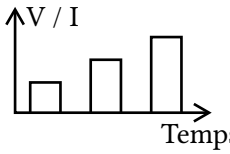
1. Une source de **tension** (comme dans l'état de l'art)
2. Une source de **courant sans** limitation locale du courant
3. Une source de **courant avec** une limitation locale du courant (en utilisant le sélecteur de la cellule 1T1R, comme pour la source de tension)

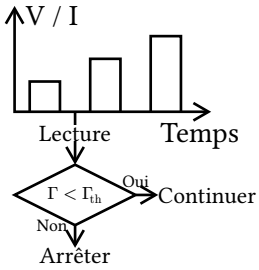
Nous explorons également deux algorithmes :

1. L'impulsion unique (opération nommée SET dans le Tableau 4.1)
2. L'écriture progressive et vérifiée (opération nommée SET Progressif et Vérifié dans le Tableau 4.1)

Tableau 4.1 – Description des opérations utilisées sur la plateforme de test du circuit *MEMORIIC*.

Nom	Description		Étapes[α]	Illustration
Lecture	Lit la conductance des cellules		`L;`	
SET	Écrit puis lit le nouvel état de conductance des cellules	Correspond à une écriture	`E_S; L;` ou `E_R; L;`	
RESET		Correspond à un effacement		

SET Progressif	Écrit les cellules petit à petit avec une série d'écritures à amplitude `A[n]` (de courant, V_+^{ext} ou V_{WL}) augmentant progressivement		`for i {`
			`  E_S(A[i]);`
			`  L;`
			`}`
				

SET Progressif et Vérifié	Comme SET Progressif, mais l'écriture sur chaque cellule s'arrête lorsque la conductance a dépassé une valeur limite `G_th` choisie par l'utilisateur		`for i {`
			`  E_S(A[i], M);`
			`  L(G_th, M);`
			`}`
				

Cyclage	Effectue `N` cycles SET/RESET		`for n in [1;N] {`
			`  E_S; L;`
			`  E_R; L;`
			`}`	

Cyclage avec SET Progressif et Vérifié	Comme Cyclage, mais avec un SET Progressif et Vérifié plutôt qu'un SET		`for n in [1;N] {`
			`  M = 0;`
			`  for i {`
			`  E_S(A[i], M);`
			`  L(G_th, M);`
			`  }`
			`  E_R; L;`
			`}`	

[α] Suite d'opérations données sous forme de pseudo code. `E_S`, respectivement `E_R`, représente un SET, respectivement un RESET, en tension ou en courant, `L` une lecture et `M = 0` la réinitialisation du masque. La présence du paramètre `M` indique que l'opération utilise le masque. Chaque opération est effectuée séquentiellement sur chaque cellule du groupe ciblé avant de passer à l'opération suivante.

Tableau 4.2 – Nomenclature des expériences de formation de filament sur le circuit *MEMORIIC*

	Source		Méthode		Limitation	
Nom	Tension	Courant	Impulsion unique	Progressif et Vérifié	Active	Inactive
V-SPCL	■		■		■	
V-PCL	■			■	■	
I-SPCL		■	■		■	
I-PCL		■		■	■	
I-P		■		■		■

Le Tableau 4.2 résume les caractéristiques des 5 méthodes essayées ainsi que leurs acronymes respectifs utilisés dans la suite[32]. La source de tension est toujours associée à la limitation de courant car sans, l'élément mémoire OxRAM serait détruit avec une forte probabilité. Ainsi nous avons deux expériences utilisant la source de tension, V-SPCL et V-PCL. Ces deux expériences sont répétées avec la source de courant, ce sont I-SPCL et I-PCL. Enfin, une cinquième expérience est conduite avec la source de courant en retirant la limitation en courant, I-P. Celle-ci n'est effectuée que pour la méthode d'écriture progressive et vérifiée car l'impulsion unique conduit à des résultats inexploitables, soit les éléments mémoires OxRAM ne sont pas formés (si le courant est trop faible) soit ils sont détruits (lorsque le courant est trop fort). I-P est la méthode la plus comparable à la méthode de FORMING utilisée dans le Chapitre III. Dans le cas de V-PCL et I-PCL, les résultats pour deux tensions V_{WL} différentes, et donc deux courants de limitations différents, sont présentés. Lorsque l'on dit qu'il n'y a pas de limitation en courant, cela signifie que V_{WL} est au potentiel de V_{dd}, le maximum possible, soit 4,6 V pour le circuit *MEMORIIC*. Dans toutes les opérations, les impulsions durent toutes 1 µs, sauf dans le cas de I-P, qui utilise des impulsions de 1 ms pour permettre au très faible courant d'atteindre des tensions permettant l'écriture, et pour encourager la formation à de plus faibles tensions pour limiter l'amplitude du pic de courant résultant de la décharge capacitive.

La Figure 4.8 montre les distributions obtenues pour chaque méthode directement après l'opération de formation du filament. De nombreuses différences apparaissent.

En commençant par les opérations à impulsion unique V-SPCL et I-SPCL, on peut remarquer que la queue de haute conductance est assez similaire entre le mode tension et le mode courant. C'est attendu puisqu'elle devrait être dominée par l'effet de la limitation de courant locale. Cependant, la queue de faible conductance est assez différente. Dans le mode tension, au sein de certains éléments mémoires OxRAM, aucun filament ne se forme (leur conductance reste essentiellement nulle), alors que toutes les cellules sont formées dans le mode courant. La formation d'un filament est un processus stochastique motivé par la présence d'un fort champ électrique au sein de l'oxyde. Dans le cas du mode tension, ce champ est principalement fixé par la source de tension. Ainsi, au sein d'un élément mémoire OxRAM, le champ électrique avant l'éventuelle formation d'un filament est fixé et n'évolue pas avec le temps. S'il est insuffisant, l'élément mémoire OxRAM peut ne pas être formé. Au contraire, dans le mode courant, comme l'élément mémoire OxRAM est initialement un bon isolant, le courant ne peut que s'accumuler dans la capacité parasite C_p. La tension augmente donc avec le temps. Ainsi, les éléments mémoires OxRAM facilement formables (car leurs défauts favorisent un champ électrique localement fort) sont formés assez vite à une tension relativement faible. Au contraire, les éléments mémoires OxRAM difficilement formables (car leur défauts favorisent un champ électrique localement faible) prendront légèrement plus de temps pour être formés car ils nécessitent une tension plus forte. De plus, dès lors que la formation du filament a lieu, la tension chute rapidement pour atteindre la valeur prévue par la loi d'Ohm. L'utilisation d'une source de courant pour la formation du filament permet donc d'implémenter une forme d'algorithme progressif de manière purement électrique.

Pour les opérations progressives et vérifiées V-PCL et I-PCL, deux conditions avec différents V_{WL} sont rapportées. Dans le cas du mode tension, on peut tout d'abord remarquer que la plupart des cellules sont formées, mais pas toutes. La vérification permet (comme attendu) de compresser la queue de faible conductance juste au-dessus de la conductance limite de vérification. On peut également voir que l'augmentation de la limite en courant (augmentation de V_{WL}) a pour effet de translater la distribution de manière assez uniforme. Cela nous indique que l'état de conductance atteint par les cellules est fortement corrélé à la limitation de courant et que la formation des filaments n'est pas

[32]De la forme A-BC avec A = V ou I pour source de tension ou source de courant, B = SP ou P pour impulsion unique ou écriture progressive et C = CL ou rien pour avec limitation du courant ou non.

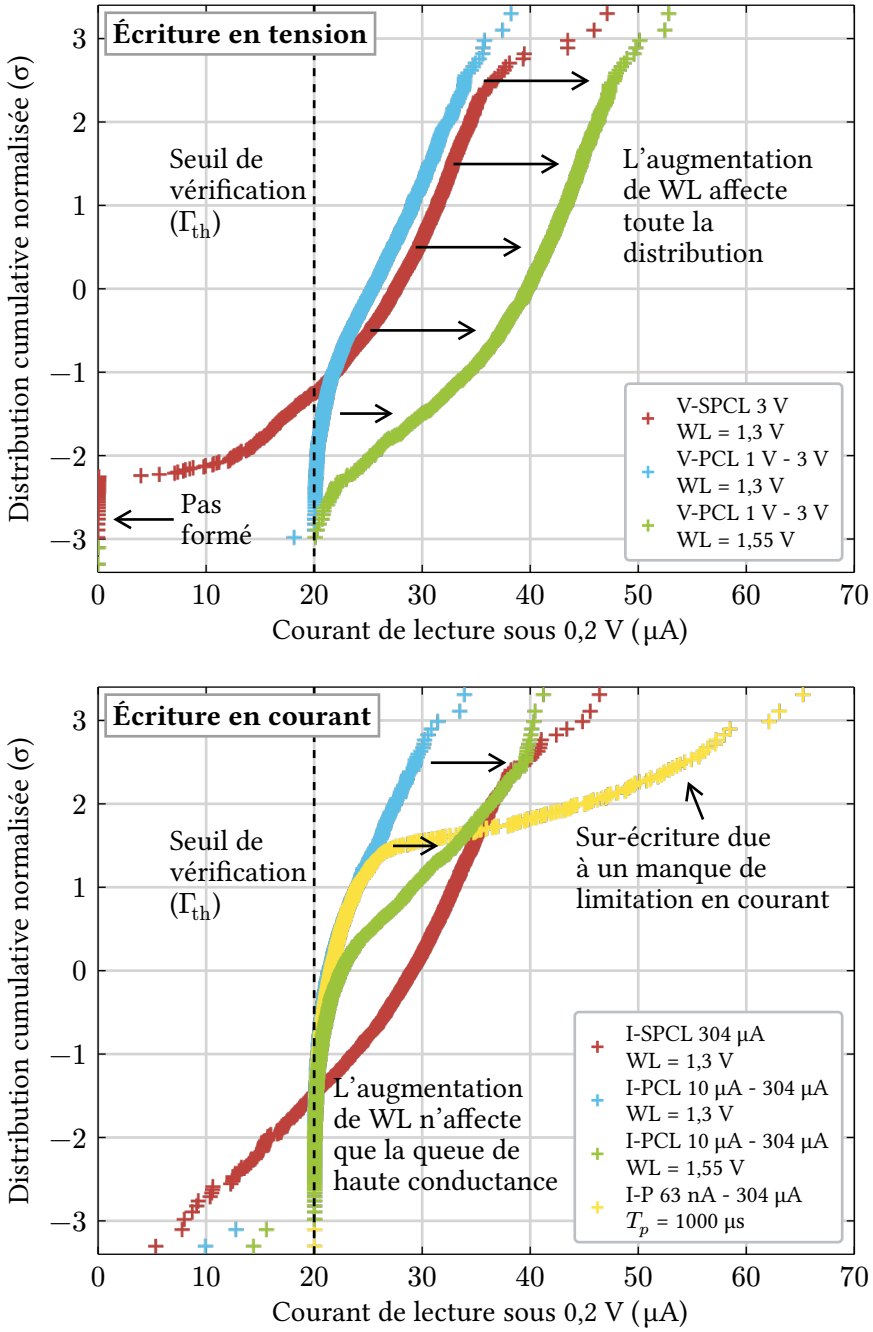


Figure 4.8 – Conductances des éléments mémoires OxRAM juste après la formation de leur filament.

progressive[33]. L'opération progressive permet d'avoir une tension adaptée à chaque élément mémoire OxRAM, permettant ainsi de former plus de cellules, sans en sur-écrire. Cela conduit à une distribution plus resserrée. Dans le cas du mode courant, toutes les cellules sont formées mais ne sont pas nécessairement au-dessus du seuil de vérification. A l'opposé du mode tension, on constate que l'augmentation du courant limite (augmentation de V_{WL}) n'affecte qu'une partie de la population, ce qui montre qu'elle n'est pas utile dans tous les cas. Cette observation est confirmée par l'expérience I-P qui n'utilise pas de limitation en courant. On constate qu'une large partie de la distribution est identique à I-PCL utilisant la limitation en courant. Cependant, on note qu'environ 10 % des cellules semblent être sur-écrites. Cette sur-écriture est problématique comme discuté ci-dessous, c'est pourquoi nous notons

[33]Du fait du fonctionnement de l'opération progressive et vérifiée, la valeur de conductance rapportée pour chaque élément mémoire OxRAM a nécessairement été atteinte en une seule étape d'écriture à partir d'une conductance sous le seuil.

Tableau 4.3 – Conditions de cyclage

	SET	RESET
Opération	Impulsion unique	
Source	Tension	
Amplitude	2 V	1,3 V
Durée	1 µs	
V_{WL}	1,3 V	4,6 V

que la limitation en courant est nécessaire lors de la formation du filament, même lorsque l'on utilise une source de courant. Dans ce cas, la limitation en courant n'a pas besoin d'être précise, elle doit simplement permettre d'absorber le pic de courant résultant de la décharge capacitive. Enfin, si l'on compare le mode tension au mode courant, outre I-P, les distributions issues du mode courant sont plus uniformes que celles en tension.

Dans l'ensemble, nous pouvons donc conclure que la méthode et les paramètres de l'opération de formation du filament impactent fortement la distribution de conductance mesurée directement après. Cependant, notre interrogation initiale porte sur les éventuels effets de cette opération sur le long terme. Pour répondre à cette question, nous effectuons des cycles d'écriture et d'effacement sur nos cellules. Pour toutes les expériences, nous utilisons les mêmes conditions de SET et RESET résumées dans le Tableau 4.3.

La Figure 4.9 montre la distribution de la conductance des cellules après 1 000 cycles. Deux distributions sont rapportées pour chaque expérience. Celle de plus faible conductance (HRS) à gauche correspond à l'état après un effacement, celle de plus forte conductance (LRS) à droite correspond à l'état après une écriture. Les différentes expériences ont des distributions très similaires, assez pour que l'on ne puisse pas voir de réelle différence dans cette représentation à l'exception de trois particularités. La première est que les cellules non formées restent non formées, elles ne sont donc pas fonctionnelles. La seconde est que les distributions HRS des deux opérations à impulsion unique semblent avoir une queue de haute conductance légèrement plus élevée que les autres opérations. Cette queue limite notre capacité à différencier les deux états et limite donc la performance de la matrice en tant que mémoire. La dernière est la queue de haute conductance de l'expérience I-P. Cette queue est la conséquence de la sur-écriture lors de l'opération de formation du filament. Elle ne semble pas réduire la fenêtre de lecture puisque elle est de haute conductance sur la distribution LRS qui est de plus haute conductance. Cependant, l'impact de l'existence de cette queue LRS sur la distribution HRS est incertain et la Figure 4.9 ne permet pas d'éclairer ce point.

En effet, dans cette représentation, le lien entre l'état LRS et HRS de chaque cellule est perdu. Comme moins de 10 % des cellules sont présents dans la queue de haute conductance de I-P, leur état HRS est complètement noyé dans la distribution. Pour mieux comprendre ce lien, nous proposons sur la Figure 4.10 une représentation des données qui conserve le lien entre les états LRS et HRS de chaque cellule. Plutôt que de trier chaque distribution indépendamment, nous groupons les conductances LRS et HRS en paire, une pour chaque cellule. Nous associons à ces paires un score correspondant à la différence entre leurs deux conductances. Puis nous les représentons par une paire de points positionnés avec en abscisse leur conductance et en ordonnée le score normal de la paire.

Dans le cas de I-PCL, nous voyons deux groupes uniformes, l'un pour HRS et l'autre pour LRS. Cette représentation permet déjà de mettre en évidence que la queue de haute conductance en HRS est due à des cellules bloquées à forte conductance qui n'ont pas été correctement RESET. Dans le cas I-P, on peut constater que deux populations de cellules semblent se distinguer. Les cellules participant à la queue de haute conductance en LRS sont associées en moyenne à des conductances HRS plus fortes.

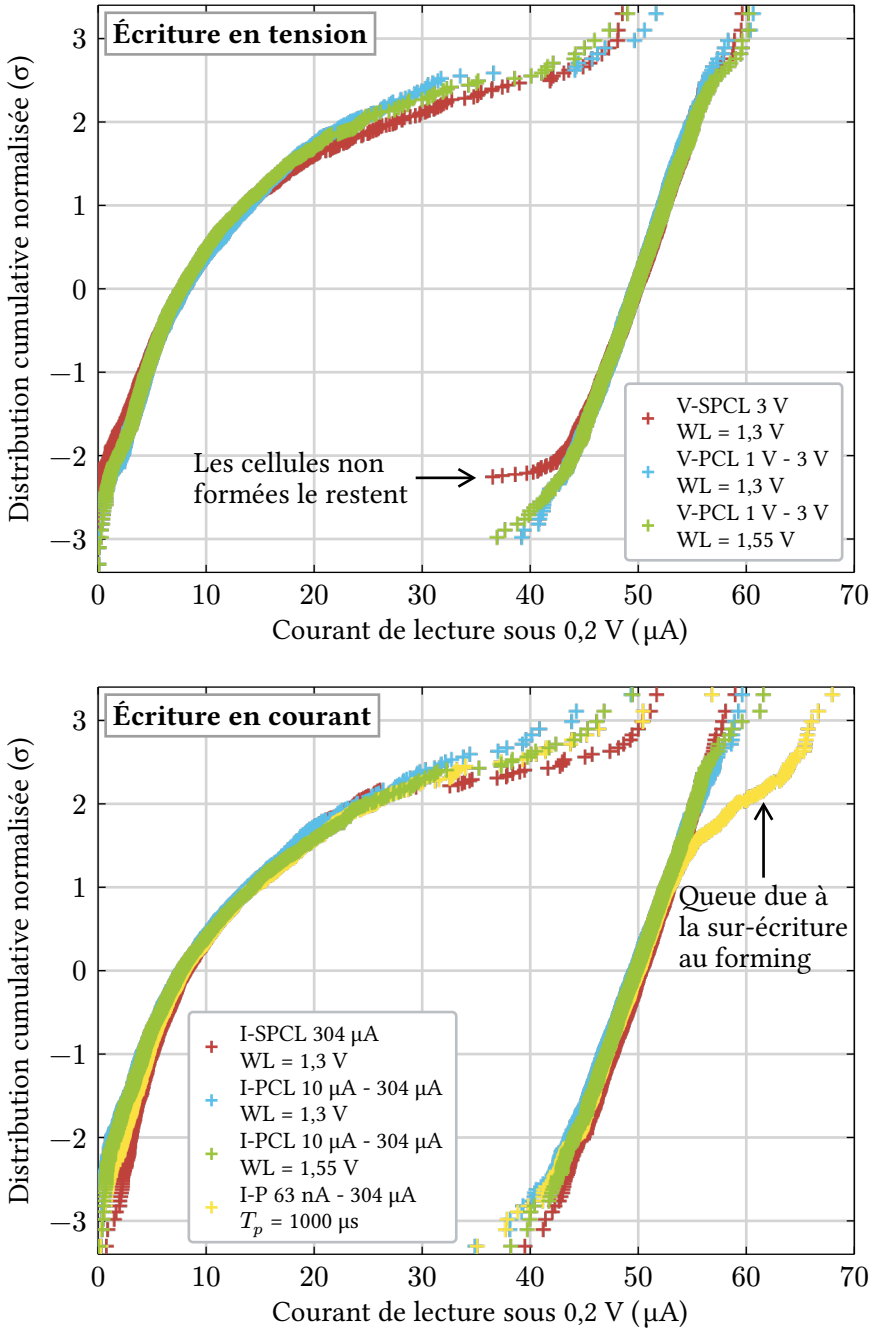


Figure 4.9 – Conductances des cellules formées après 1 000 cycles. Deux distributions sont dessinées pour chaque expérience. Celle de gauche correspond à l'état de plus faible conductance après un RESET, celle de droite correspond au SET.

La distribution de leur marge de lecture est distribuée de même manière que celle des autres cellules. Ainsi, c'est comme si l'on avait deux populations de cellules partageant la même matrice, avec des propriétés similaires, mais opérant à deux conductances moyennes différentes. Ceci est un problème car cela diminue l'uniformité au sein de la matrice. En effet, pour effectuer la lecture, nous avons besoin de choisir une référence permettant de différencier les deux niveaux. Or ce choix ne pourra pas satisfaire les deux populations de manière optimale. Les performances seront donc réduites. En conséquence, nous notons que la queue LRS de haute conductance est un problème pour la performance de la matrice : I-P n'est pas satisfaisant.

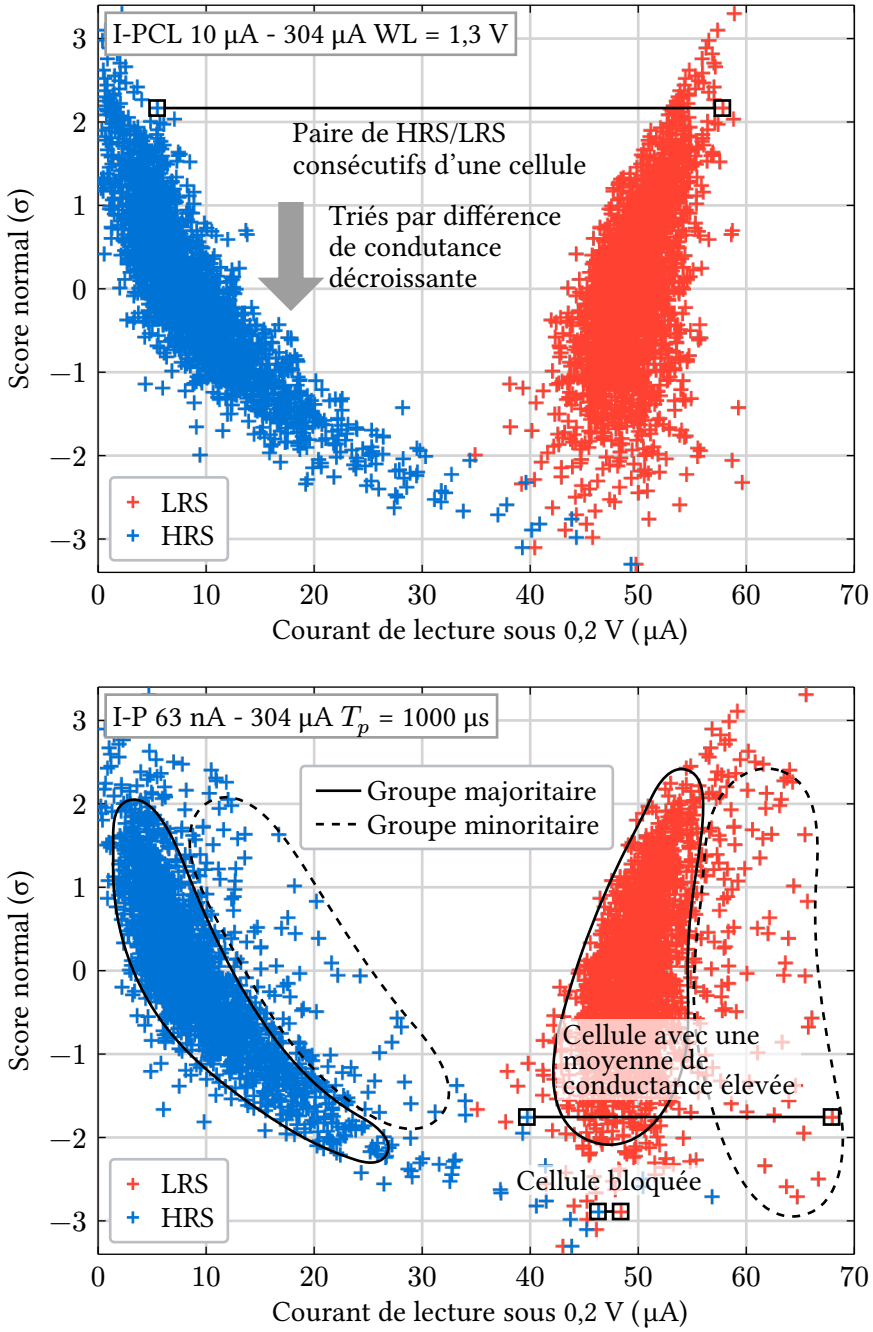


Figure 4.10 – Tracé alternatif des données de la Figure 4.9. Ici les points des deux distributions sont appairés par leur adresse d'origine. Un score proportionnel à la différence de conductance entre les deux membres de la paire leur est associé. Leur ordonnée est leur score normal et leur abscisse leur conductance. Les deux points de chaque paire sont dessinés, l'un en bleu (HRS) et l'autre en rouge (LRS).

Nous pouvons donc conclure que, dans la mesure du raisonnable, les paramètres de l'opération de formation du filament ont peu d'impact sur le comportement des éléments mémoires OxRAM comme expliqué par [73]. Cependant, dans des cas plus extrêmes, par exemple ici I-P, un impact non négligeable est observé, comme dans le cas du FORMING à chaud de [86]. Dans notre cas, tant que toutes les cellules sont formées et pas sur-écrites, les performances après cyclage semblent être similaires. Pour la suite de nos expériences, nous sélectionnons les paramètres de l'expérience I-PCL avec $V_{\mathrm{WL}} = 1{,}3$ V comme conditions de FORMING.

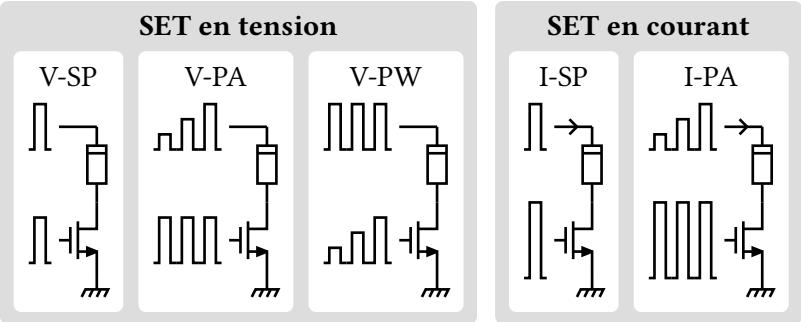


Figure 4.11 – Stratégies de SET utilisées

Tableau 4.4 – Paramètres utilisés pour les stratégies de SET

Nom	Amplitude	Durée (µs)	V_{WL} (V)
I-SP	200 µA	1	4,6
I-PA	20–304 µA [α]		
V-SP	2 V		1,4
V-PA	0,8–2 V [β]		
V-PW	1,3 V		1–1,8 [γ]

[α] L'incrément de courant démarre à 1 µA et est doublé à chaque fois qu'une puissance de deux est atteinte.
[β] L'incrément d'amplitude en tension est 150 mV
[γ] L'incrément de V_{WL} est 100 mV

Ces résultats ont fait l'objet d'une communication à un congrès, l'International Memory Workshop (IMW) à Séoul en 2024. (DOI: 10.1109/IMW59701.2024.10536941)

IV.4.2 Écriture

Après avoir déterminé une condition de formation du filament viable dans la partie précédente, nous étudions l'opération d'écriture SET, c'est-à-dire le passage de l'élément mémoire OxRAM d'une faible conductance à une forte conductance. L'objectif est de juger de l'intérêt de la technique d'écriture en courant par rapport à celle en tension.

La Figure 4.11 illustre les stratégies d'écriture étudiées ici. Pour le mode courant, nous retrouvons essentiellement les deux stratégies étudiées pour la formation du filament, c'est-à-dire l'impulsion unique (I-SP) ainsi que l'écriture progressive vérifiée (I-PA). La différence réside dans le fait qu'ici, nous n'étudions le mode courant que sans limitation en courant. Elle est en effet moins nécessaire qu'au Forming car la plus faible variation de conductance génère un pic de courant plus faible. I-SP est semblable à la méthode utilisée dans le Chapitre III.

Du côté du mode tension, on retrouve l'impulsion unique (V-SP), mais l'on a cette fois deux versions d'écritures progressives vérifiées. La première module l'amplitude de V_+ (V-PA). La seconde module V_{WL}, et donc la valeur de limitation du courant (V-PW). Le Tableau 4.4 donne les principaux paramètres choisis pour chaque stratégie. La durée de chaque impulsion est 1 µs. Elle est identique pour toutes les stratégies. La tension V_{WL} utilisée en mode tension est légèrement plus forte que celle utilisée pour la formation du filament. Le courant de seuil utilisé pour la vérification est 40 µA pour V-PA. Deux expériences sont effectuées pour V-PW et I-PA, l'une avec le seuil 40 µA et l'autre 30 µA. Dans toutes ces expériences, l'effacement (RESET) est effectué avec une écriture à impulsion unique avec $V_+ = 1{,}3$ V, une durée de 1 µs et $V_{WL} = 4{,}6$ V.

La Figure 4.12 rapporte les distributions obtenues après 2 000 cycles.

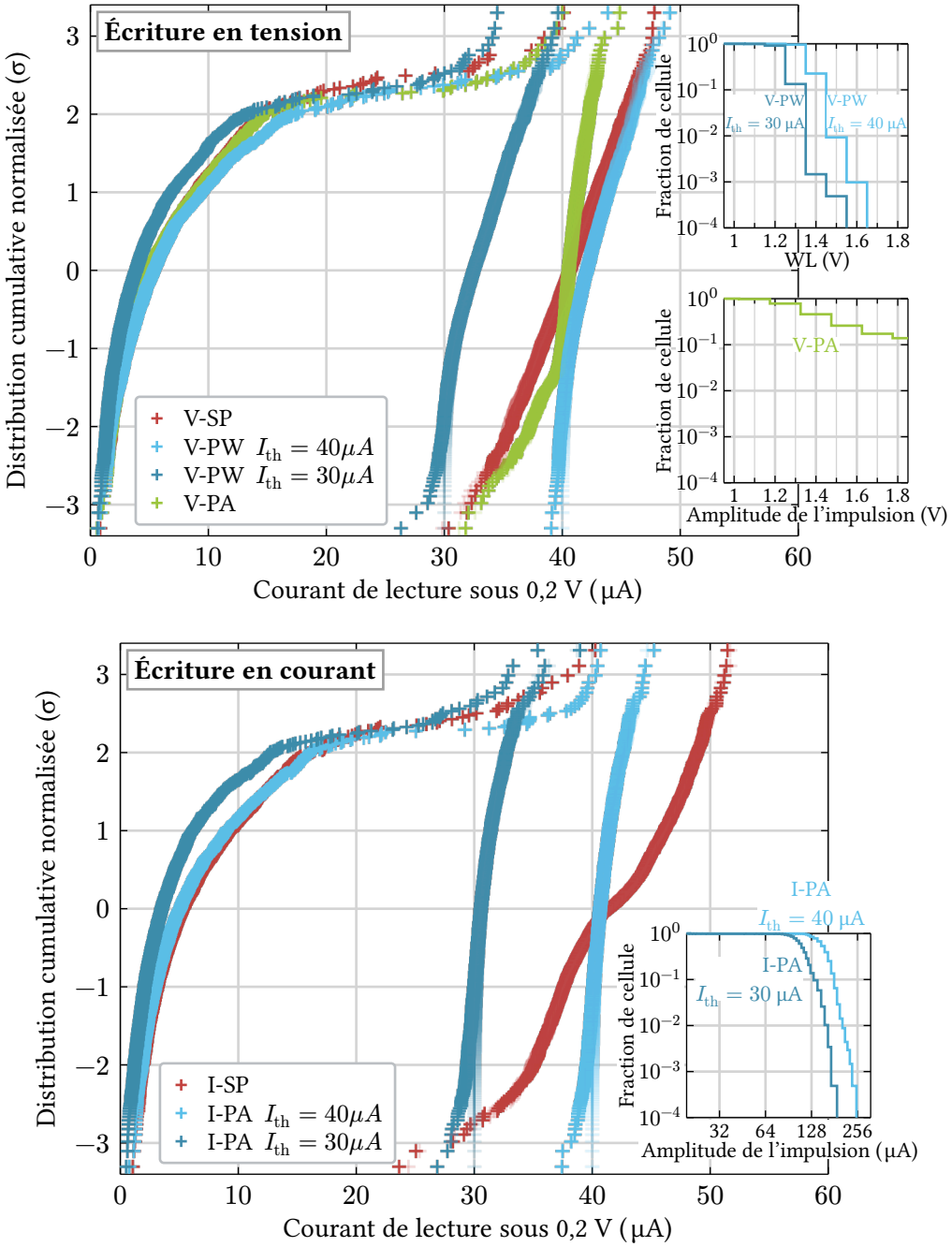


Figure 4.12 – Distributions cumulées des courants de lecture après 2 000 cycles. Les distributions HRS (gauche) et LRS (droite) sont toutes deux représentées. Les points de couleur transparente, respectivement pleine, correspondent au courant mesuré immédiatement après l'écriture, respectivement après une période de relaxation de 10 s. Les encarts correspondent aux histogrammes de la proportion d'éléments mémoires OxRAM ayant effectivement subi l'impulsion en abscisse (le complément ayant déjà atteint le seuil de vérification) pour les méthodes utilisant la vérification.

En commençant par le mode tension, V-SP présente la distribution LRS la plus large. En particulier la queue de faible conductance contraint la fenêtre de lecture, d'où l'intérêt des méthodes d'écriture vérifiées. V-PA ne permet pas d'écrire tous les éléments mémoires OxRAM au-dessus du seuil de vérification. Cela aurait peut-être été possible avec une plage de tension plus élevée. Cependant, les histogrammes en encart nous permettent de voir que la modulation de V_+ a beaucoup moins d'effet

que la modulation de V_{WL} comme noté par [36] [87]. Les distributions HRS sont très proches. Seule celles de V-PW avec $I_{\mathrm{TH}} = 30\,\mu\mathrm{A}$ se démarquent avec une queue haute conductance légèrement moins prononcée entre 1σ et 2σ.

La même remarque peut être faite pour le mode courant. Les distributions HRS sont toutes similaires sauf celle de I-PA avec $I_{\mathrm{TH}} = 30\,\mu\mathrm{A}$. Les distributions LRS en I-PA sont plus resserrées que leur homologue en tension V-PW. C'est *a priori* un point positif. Comme on peut le voir dans les histogrammes en encart, I-PA est beaucoup plus progressif que V-PW. Il est donc difficile de dire si cette différence est causée par le type de source utilisée, ou par la différence de granularité de programmation.

En comparant les données obtenues avec une lecture immédiatement après écriture avec celles obtenues après une attente de 10 s (points transparents pour les premiers et pleins pour les seconds sur la Figure 4.12), on constate que la relaxation affecte quasi-exclusivement les distributions issues des écritures vérifiées proches du courant seuil. Cependant, cela ne signifie pas que la vérification amplifie la relaxation, un effet statistique est suffisant pour expliquer ces observations.

De façon imagée, on peut imaginer une foule contenue dans un cercle de 100 m de diamètre (un stade par exemple). Si chaque individu se déplaçait aléatoirement dans un rayon de 5 m autour de leur position initiale, la foule resterait semblable à ce qu'elle était avant le déplacement. Son rayon serait au plus augmenté de 5 m. Par contre, si chaque individu venait à se déplacer de 1 km, la foule serait grandement modifiée et occuperait de façon assez uniforme un disque d'un peu plus de 1 km de rayon. Ainsi, il est attendu que la relaxation affecte principalement les parties de la distribution qui ont une dispersion faible devant le déplacement de conductance, de l'ordre de 10 µS dans notre cas, qu'elle engendre. Proche des queues de distribution, il y a également un effet discret qui se produit. Selon la même analogie, si l'on se place près du bord de la foule, ce bord devrait en théorie se déplacer de 5 m. Cependant, si le nombre de personnes initialement proches du bord est faible, il est possible qu'aucune ne se déplace assez loin et dans la bonne direction pour matérialiser ce nouveau bord théorique. Plus la foule est dense, moins ce cas est probable. Donc, plus la distribution est dense, proche d'un bord de son support, plus la distribution empirique relaxée a de chance d'exhiber un de ses points extrêmes. Comme la vérification crée une grande accumulation de points près du seuil, c'est à cet endroit que la relaxation est la plus observable. A la fois car la dispersion y est faible, mais également car beaucoup de points sont candidats pour avoir un déplacement extrême vers le bas et être observé dans la queue de la distribution relaxée.

La distribution LRS de I-SP présente une irrégularité singulière. Elle semble être coupée en deux lobes normaux de même variance, mais de moyenne différente. Cela ressemble à la séparation de population observée lors de la formation du filament sans limitation de courant. Comme pour l'étude de la formation du filament, regarder les données sous une autre représentation peut s'avérer utile. La Figure 4.13 montre les données sous la même forme que la Figure 4.10, à la différence que le point HRS choisi pour chaque adresse est celui de plus forte conductance parmi les 1 000 derniers cycles, le point LRS correspondant est l'état juste avant d'atteindre le point HRS choisi. Cela nous permet immédiatement de confirmer que l'on retrouve bien une séparation de population dans le cas d'I-SP. Il est à noter que I-SP est en théorie plus sensible au manque de limitation en courant que I-PA. En effet, comme l'opération n'est pas progressive, un fort courant est directement appliqué aux éléments mémoires OxRAM. Le courant élevé conduit à une élévation rapide de la tension. Or, plus le gradient de tension est fort, plus la tension moyenne à laquelle l'écriture se produit est élevée [16]. Ainsi, les pics de courant causés par la capacité parasite C_p devraient être plus forts dans le cas de I-SP que dans celui de I-PA. Le phénomène d'écriture aidée par le pic de courant observé et décrit au Chapitre III se révèle donc ne pas fonctionner de manière homogène sur les différentes cellules. Nous pouvons donc constater l'importance de l'expérimentation sur un nombre important de cellules. Les conclusions du précédent chapitre ne sont pas nécessairement erronées, mais le doute que nous avions émis sur

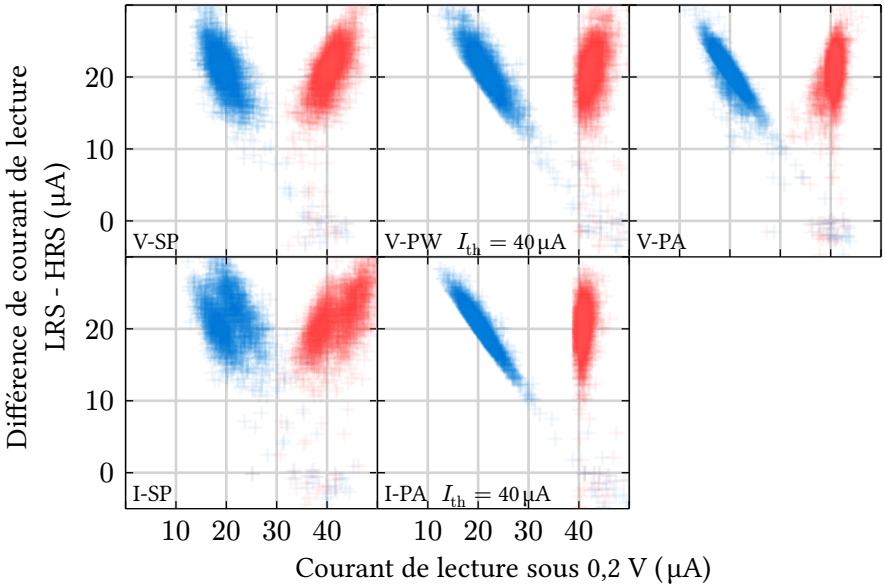


Figure 4.13 – Représentation alternative, similaire à la Figure 4.10, des données de Figure 4.12. Contrairement à la Figure 4.10, la valeur de conductance HRS la plus élevée de chaque adresse sur les 1 000 derniers cycles est appairée avec la conductance LRS qui l'a précédée.

leur validité générale, du fait de la faible statistique à laquelle nous avions accès ainsi qu'à l'étape de sélection quelque peu arbitraire des cellules après leur formation, se révèle être justifié.

Avant de continuer, nous tenons à faire remarquer que la représentation des données sous la forme présentée en Figure 4.13 peut facilement être sur-interprétée. En effet, de par la construction de cette représentation, une forte structure existe. Comme chaque point correspond en réalité à une paire, et comme elles sont ordonnées selon leur différence de courant, la forme du nuage HRS est fortement liée à celui de LRS. Cette relation est particulièrement notable sur les données de I-PA où la distribution LRS très resserrée autour de 40 µA conduit à un nuage HRS suivant la ligne définie par $y = -x + 40$ avec y l'ordonnée et x l'abscisse en µA. On ne peut donc pas conclure que la distribution LRS a une forte incidence sur la distribution HRS, visuellement, ce n'est qu'une conséquence de la structure de la représentation.

Cependant, on peut identifier certaines choses. Sur les données de I-SP, on voit clairement se dessiner les deux populations. L'une est centrée sur une conductance LRS plus élevée que l'autre, et dans un même temps, on voit que la fenêtre de lecture est également plus large en moyenne (le nuage est plus haut sur la représentation graphique). Cela rejoint la corrélation généralement observée entre valeur moyenne de la conductance LRS et valeur moyenne de la fenêtre [84].

Si l'on imagine maintenant devoir placer un courant de lecture I_R marquant la limite entre les niveaux que l'on considère HRS ou LRS, on peut identifier ce qui pourrait contraindre ce choix *a posteriori* et nous empêcher de réduire le taux d'erreur.

En premier lieu, les points vers le bas de la représentation sont évidement problématiques. Ces points correspondent à des cellules bloquées à forte conductance. En effet, bien que la valeur de HRS la plus élevée de chaque cellule dans les 1 000 derniers cycles soit représentée, seules quelques dizaines de paires de points sont dans cette zone. Ainsi, on est certain que la vaste majorité des cellules ne se retrouve jamais durant ces 1 000 cycles dans cette situation. Après vérification manuelle de l'historique d'élément dans cette zone, on peut affirmer que ce sont bien des éléments bloqués. On peut donc considérer que ces éléments mémoires OxRAM sont défectueux et que l'on ne pourra pas s'en servir.

Ensuite, on remarque que la queue faible conductance de la distribution LRS dans le cas de V-PA est problématique. En effet, elle nous force à réduire I_R, ce qui augmentera nécessairement les erreurs dues à la queue haute conductance en HRS.

Au contraire, la queue haute conductance en LRS de V-PW par rapport à I-PA ne semble pas poser de problème concernant I_R. En effet, elle a tendance à « diffuser » les point HRS plus proches de I_R mais ne les contraint pas directement. Cependant, elle pourrait causer d'autres problèmes sur des performances non observées ici. On peut donc considérer qu'il est tout de même préférable de la réduire.

Ainsi, d'après ces considérations, I-PA semble être la meilleure stratégie, avec V-PW pas loin derrière. Ensuite viennent V-PA et V-SP avec un avantage pour V-PA qui a une distribution LRS plus compacte. Enfin, I-PS apparaît comme la stratégie la moins satisfaisante du fait de la mauvaise homogénéité qu'elle provoque.

Cette analyse est pour l'instant purement qualitative. Nous effectuons par la suite l'équivalent de façon quantitative. Pour cela, nous cherchons I_R minimisant le taux d'erreur. Nous opérons par approximations successives en affinant notre estimation à différente échelle. Nous commençons sur un intervalle large. Nous calculons le taux d'erreur pour une certain nombre de valeurs de I_R uniformément réparties sur l'intervalle. Le taux considéré est le nombre d'erreurs (un LRS lu comme un HRS et inversement) par rapport au nombre d'erreurs possibles. La valeur de I_R minimisant ce taux est sélectionnée. Puis l'opération est répétée sur un intervalle plus petit centré sur la valeur sélectionnée. Le résultat n'est pas nécessairement la valeur optimale, mais de part la régularité induite par le fait que nos fonctions sont des distributions cumulées, la valeur de I_R trouvée ne devrait pas provoquer plus de quelques erreurs supplémentaires que l'optimum.

A l'aide de cette valeur de I_R, nous calculons ensuite le nombre d'erreurs sur chaque adresse au cours des 1 000 derniers cycles. Les résultats obtenus sont représentés sur la Figure 4.14. Les cellules identifiées plus tôt comme bloquées constituent l'intégralité du bloc rouge (taux d'erreur supérieur à 10 %).

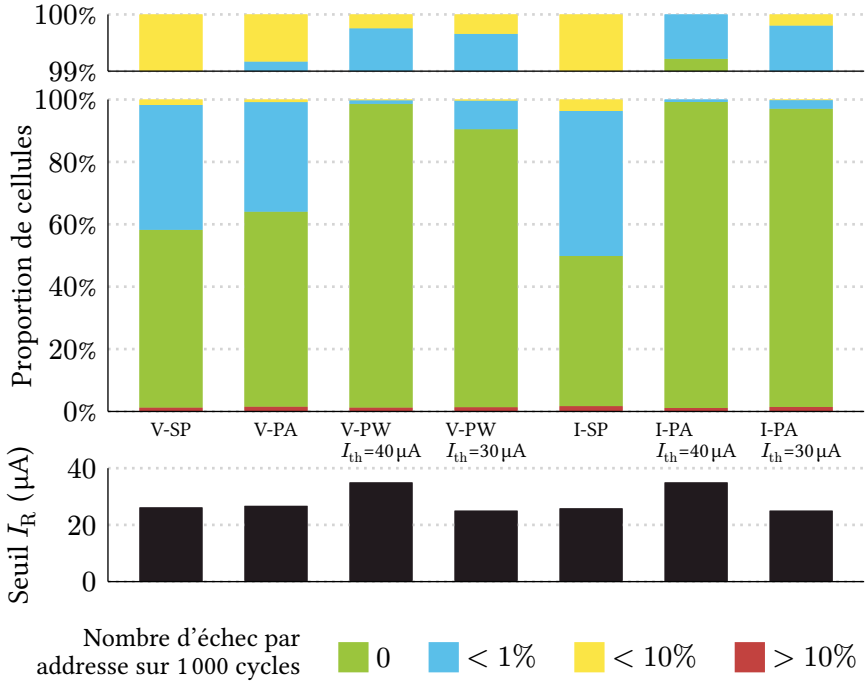


Figure 4.14 – Histogramme de la population de cellules dans une certaine tranche de taux d'erreur en fonction de la méthode d'écriture utilisée (en haut), et histogramme I_R estimé comme optimum et utilisé pour le calcul du taux d'erreur (en bas).

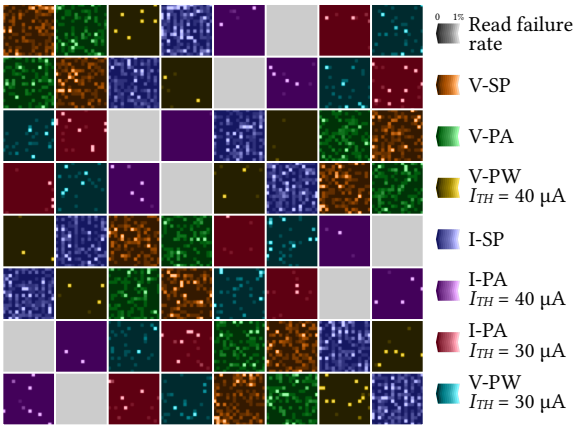


Figure 4.15 – Cartographie de la matrice. La nuance indique la stratégie utilisée dans la région, et la luminosité indique le taux d'erreur.

On peut constater qu'en classant les méthodes par leur taux d'erreur, on retrouve notre précédent classement établi de manière qualitative. Les deux expériences avec $I_{\text{th}} = 30\,\mu\text{A}$ présentent légèrement plus d'erreurs que leurs homologues avec $I_{\text{th}} = 40\,\mu\text{A}$, ce qui est attendu. En effet, ces dernières peuvent avoir un I_{R} plus fort ayant donc moins de chance de provoquer une erreur en confondant un état HRS en un état LRS. Cependant on peut voir que les stratégies V-PW et I-PA avec $I_{\text{th}} = 30\,\mu\text{A}$ ont de meilleures performances que les autres stratégies malgré leur I_{th} comparable voir inférieur à ceux des autres. Leur avantage est donc nécessairement pris sur une meilleure distribution HRS. Par conséquent, on peut conclure que ces stratégies ont un effet observable sur l'état HRS bien qu'elles ne concernent que l'écriture SET (passage de HRS vers LRS).

La Figure 4.15 présente l'information du nombre d'erreurs par adresse sous forme de cartographie. L'intérêt de cette représentation est de s'assurer que la distribution des erreurs n'a pas de structure. En effet, on s'attend à trouver des erreurs distribuées aléatoirement lorsqu'elles proviennent d'un défaut de l'élément mémoire OxRAM. Si elle est structurée, cela peut indiquer un défaut dans le circuit ou dans la conception de la matrice. Tout semble assez bien distribué, sauf pour I-SP. Visuellement, il semble apparaître des bandes verticales, comme si les colonnes paires avaient plus de probabilité de contenir une erreur. On rappel que I-SP est la stratégie menant à la séparation en deux populations. On s'attend à obtenir la majorité des erreurs chez la population de plus forte conductance moyenne. Donc cette cartographie pourrait indiquer que les deux sous-populations causées par I-SP ne sont pas distribuées aléatoirement. Cependant, l'architecture de la périphérie de la matrice ne paraît pas, à première vue, induire une telle structure. Une évaluation plus poussée serait nécessaire pour conclure, cependant cette direction n'a pas été approfondie.

En conclusion, I-PA semble être la meilleure stratégie parmi celles étudiées. Cependant, V-PW présente des caractéristiques similaires et l'on rappelle que la différence concernant la queue de haute conductance en LRS pourrait provenir de la différence de granularité de leur progressivité. D'un point de vue de l'intégration, la solution I-PA semble plus simple à intégrer. En effet, il est plus facile de moduler et de recopier un courant qu'une tension. En conséquence, pour réaliser I-PA, une seule source de courant réglable serait nécessaire, et l'écriture en parallèle pourrait être réalisée à l'aide d'une copie de ce courant. Alors que dans le cas de V-PW, soit une source de tension réglable doit être disponible pour chaque cellule à écrire en parallèle, soit il est nécessaire de copier cette tension ou de dimensionner la source pour pouvoir fournir du courant à plusieurs cellules en même temps (ce qui pourrait induire des effets de couplage avec une variation de la tension sur l'ensemble des cellules lorsqu'une cellule SET). De plus I-PA ne nécessite pas de limiteur de courant ce qui fait un élément de moins à intégrer et gérer.

Ces résultats ont fait l'objet d'une communication à un congrès, le European Solid-State Electronics Research Conference (ESSERC). (DOI: DOI: 10.1109/ESSERC62670.2024.10719434)

IV.5 Conclusion

La conception du circuit *MEMORIIC*, spécialement réalisée pour nos expériences, nous a permis de dépasser les limitations que nous avions rencontrées lors de l'exploitation du circuit *WT-PAOLA*. Les points clés ont été l'intégration d'une matrice de cellules d'éléments mémoires OxRAM permettant l'obtention de plus larges statistiques, et la possibilité de limiter le courant permettant l'utilisation des méthodes d'écriture en tension pour effectuer une comparaison directe avec les méthodes d'écriture en courant.

Le développement de la plateforme de tests a également été un point clé dans l'obtention des résultats présentés. Le choix de maximiser l'autonomie de la plateforme nous a permis de tester, de faire évoluer la plateforme et de l'exploiter plus et plus rapidement, avec également un confort d'utilisation accru. Grâce à la disponibilité de l'historique complet des actions menées sur chaque cellule, nous avons pu effectuer des analyses complémentaires *a posteriori*, comme la vérification de l'historique des cellules qui étaient supposées bloquées lors du cyclage en SET. De plus, le maintien de cet historique n'a pas été une difficulté grâce à sa considération dès le début du développement de la plateforme. Les autres choix d'architectures soulignés tels que limiter le logiciel du micro-contrôleur à un simple interpréteur de commandes élémentaires, et développer un logiciel monolithique pour l'ordinateur avec une interface utilisateur riche et dynamique n'ont pas été regrettés.

Tous ces développements nous ont permis d'effectuer les expériences que nous souhaitions faire concernant l'écriture en courant. L'accent a été mis sur la comparaison avec l'écriture en tension sur la même puce, dans l'objectif de comparer ces écritures avec le moins de biais possible. A propos du FORMING, nous avons vu que les paramètres et la méthode utilisée ont assez peu d'importance après un millier de cycles. Cependant, nous avons également remarqué qu'un FORMING plus extrême pouvait avoir des conséquences même après cyclage. Le cas du SET nous a montré que les écritures progressives et vérifiées, aussi bien en tension qu'en courant, proposaient le plus faible taux d'erreur, avec un avantage pour l'écriture en courant qui génère une distribution LRS plus homogène sur la matrice.

Bibliographie

[16] G. Sassine, C. Cagli, J.-F. Nodin, G. Molas, et E. Nowak, « Novel Computing Method for Short Programming Time and Low Energy Consumption in HfO2 Based RRAM Arrays », *IEEE Journal of the Electron Devices Society*, vol. 6, p. 696-702, 2018, doi: 10.1109/JEDS.2018.2830999.

[36] V. Milo *et al.*, « Accurate Program/Verify Schemes of Resistive Switching Memory (RRAM) for In-Memory Neural Network Circuits », *IEEE Transactions on Electron Devices*, vol. 68, nº 8, p. 3832-3837, août 2021, doi: 10.1109/TED.2021.3089995.

[73] C.-Y. Wu *et al.*, « Emerging Memory RRAM Embedded in 12FFC FinFET Technology for industrial Applications », in *2 023 International Electron Devices Meeting (IEDM)*, déc. 2023, p. 1-4. doi: 10.1109/IEDM45741.2023.10413775.

[84] C. Cagli *et al.*, « Study of the Energy Consumption Optimization on RRAM Memory Array for SCM Applications », in *2 017 IEEE International Memory Workshop (IMW)*, 2017, p. 1-4. doi: 10.1109/IMW.2017.7939106.

[86] X. Xu *et al.*, « 40× Retention Improvement by Eliminating Resistance Relaxation with High Temperature Forming in 28 nm RRAM Chip », in *2 018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 20. doi: 10.1109/IEDM.2018.8614593.

[87] E. Perez, M. K. Mahadevaiah, E. P.-B. Quesada, et C. Wenger, « Variability and Energy Consumption Tradeoffs in Multilevel Programming of RRAM Arrays », *IEEE Transactions on Electron Devices*, vol. 68, nº 6, p. 2693-2698, juin 2021, doi: 10.1109/TED.2021.3072868.

Chapitre V : Écriture capacitive

V.1 Introduction

Dans les chapitres précédents, nous avons discuté de l'écriture des éléments mémoires OxRAM en utilisant une source de tension, puis une source de courant. Nous avons vu que le comportement observé était différent, ce qui motive la question suivante, « Pourrait-on utiliser autre chose ? ». Dans ce chapitre, nous explorerons l'idée d'utiliser une capacité préchargée comme source électrique.

Une capacité préchargée ainsi utilisée comme source a pour principal attribut d'être à énergie fixe et finie. Ainsi, contrairement à la source de tension ou à la source de courant, ici, la quantité d'énergie investie dans l'écriture est limitée à une valeur connue *a priori*. Cela produit donc une forme de « terminaison d'écriture » en limitant l'énergie dissipée au sein de l'élément mémoire OxRAM. Le contrôle de l'énergie pourrait également permettre d'augmenter la vitesse d'écriture. En effet, la tension de départ de l'écriture devrait pouvoir être augmentée sans pour autant augmenter la sur-écriture comme cela serait le cas avec la source de tension.

Cependant, seule l'énergie totale est contrôlée. Comme l'énergie de SET est variable, l'énergie dissipée après le SET l'est également. Dans le contexte d'une matrice d'éléments mémoires OxRAM, la localisation de l'élément peut également influencer sur la proportion d'énergie effectivement dissipée sur l'élément, le reste l'étant sur des parasites. La quantité d'énergie disponible pour l'écriture doit être adaptée au cas en requérant le plus, ce qui pourrait affecter négativement les éléments en requérant moins en les sur-écrivant. Le contrôle de l'énergie disponible vient également au prix du contrôle de la tension dont l'évolution dans le temps est maintenant liée à la conductance de l'élément mémoire OxRAM.

Dans ce chapitre, nous explorons le fonctionnement théorique d'une écriture capacitive ainsi qu'une méthode permettant de réduire la capacité nécessaire en contre partie d'une complexité du système de contrôle accrue. Puis nous discuterons du circuit intégré *CAPA-WR* développé pour permettre l'étude expérimentale de cette méthode. En effet, l'expérimentation est rendu plus complexe par le rôle important que joue l'environnement de l'élément mémoire OxRAM. Pour obtenir des résultats représentatifs pour l'utilisation du dispositif dans une mémoire, il nous est nécessaire de co-intégrer la source capacitive avec les éléments mémoires OxRAM. La simulation n'est également pas une option, car il n'existe pas, à notre connaissance, un modèle prédictif du comportement de l'élément mémoire OxRAM lors de son écriture.

Cette étude s'inscrit dans la poursuite d'un travail commencé par Paola Trotti lors de sa thèse. Nous utiliserons en particulier les résultats de [79].

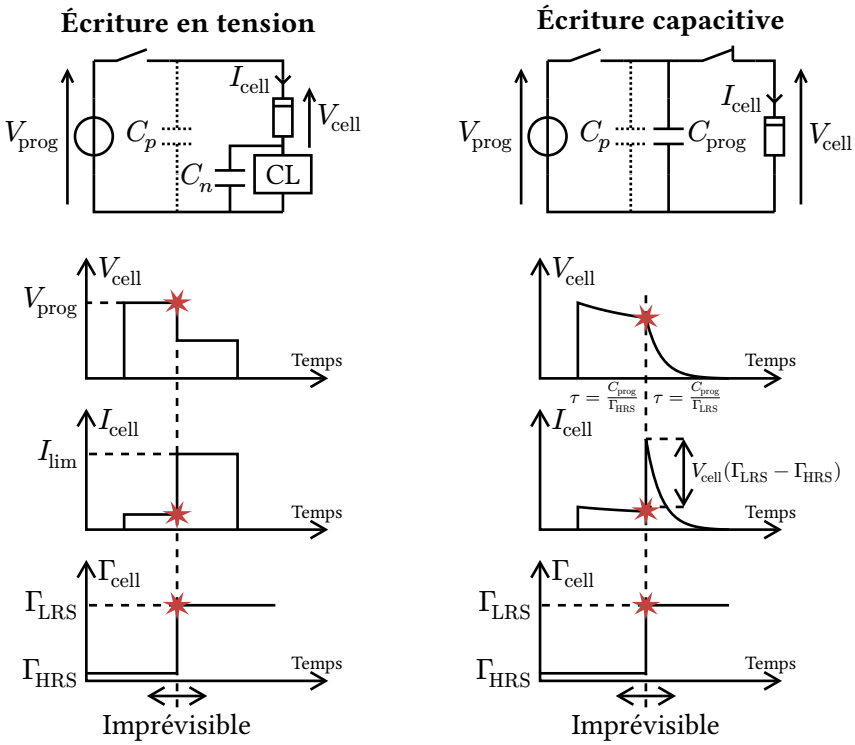


Figure 5.1 – Illustration de l'évolution de la tension sur l'élément mémoire OxRAM et du courant dans l'élément mémoire OxRAM pour une écriture avec une source de tension et pour une écriture avec une source capacitive.

Ce chapitre ne contient pas de section expérimentale. Ce projet ayant démarré seulement un peu plus d'un an avant la fin de cette thèse, le circuit *CAPA-WR* n'a pas eu l'opportunité d'être testé avant la fin de celle-ci. La fin de sa fabrication est attendue fin 2024, soit à la fin de cette thèse.

V.2 Prédiction théorique, motivations et risques

V.2.1 Principe de l'écriture capacitive

La Figure 5.1 illustre la différence entre une écriture utilisant une source de tension, et celle étudiée dans ce chapitre qui utilise une source capacitive.

Dans le cas de l'utilisation d'une source de tension, on identifie deux phases durant l'impulsion, une avant le SET et une après le SET. Avant le SET, la tension sur l'élément mémoire OxRAM est à peu près égale à la tension de la source, en effet, le courant est alors faible car l'élément mémoire OxRAM est en HRS et le limiteur de courant (CL sur la Figure 5.1) agit comme une faible résistance. Après le SET, l'élément mémoire OxRAM est en LRS, si la tension sur l'élément mémoire OxRAM était la même que celle dans la précédente phase, le courant dépasserait le courant de saturation du limiteur de courant. Celui-ci limite donc le courant, ce qui a pour effet de diminuer la tension sur l'élément mémoire OxRAM. Dans cette seconde phase, le courant dans l'élément mémoire OxRAM est donc égal à la limite en courant imposée, et la tension inférieure à celle de la source, assez pour que la réaction électrochimique au sein de l'élément mémoire OxRAM soit fortement ralentie et essentiellement arrêtée à l'échelle de temps mise en jeu.

Dans le cas de l'utilisation d'une capacité pré-chargée en tant que source, on peut également identifier les deux même phases : avant et après le SET. Avant le SET, la tension sur l'élément mémoire OxRAM diminue lentement car il est en HRS, ce qui conduit à un faible courant de décharge de la capacité. Dès que le SET se produit, le courant augmente soudainement en réaction à la chute rapide de la résistivité de l'élément mémoire OxRAM. La vitesse de décharge de la capacité augmente et rapidement la tension

diminue jusqu'à s'annuler. La seconde phase ressemble au pic de courant observé lors de l'écriture avec une source de courant, sauf que lorsque l'équilibre est atteint, la tension et le courant sont nuls.

Il est intéressant de remarquer que le temps nécessaire avant que le SET se produise est aléatoire et distribué selon une loi log-normale [16]. Lors de l'utilisation d'une source de tension, on constate une forte consommation après le SET. L'utilisation d'une impulsion plus courte augmente les chances de ne pas écrire correctement l'élément mémoire OxRAM. L'utilisation d'une impulsion plus longue conduit à une sur-consommation et à une sur-écriture de l'élément plus importantes. La forme de distribution log-normale, qui concentre sa densité au niveau des plus faibles valeurs, renforce d'autant plus l'effet néfaste de se compromis. Plus l'on applique une impulsion longue pour maximiser les chances d'écrire correctement l'élément mémoire OxRAM, plus on augmente la proportion du temps passé en moyenne dans la seconde phase de l'écriture. Ainsi, l'écriture en tension, sans un système additionnel, provoque un large gaspillage d'énergie et conduit à une potentielle sur-écriture des éléments mémoires OxRAM. Dans le cas de l'utilisation de la capacité comme source, l'énergie totale dépensée est indépendante du moment où le SET a lieu. L'étude [79] de Paola Trotti détaille ce point de manière quantitative. Bien que des techniques telles que la « terminaison d'écriture » [25] permettent de limiter ce problème lors de l'utilisation d'une source de tension, l'utilisation d'une capacité pré-chargée comme source le résout sans nécessiter aucun système ni aucune stratégie additionnels.

Le pic de courant suivant le SET rappelle celui observé lors de l'utilisation d'une source de courant. Ils ont tous les deux la même origine, la décharge rapide d'une capacité en parallèle de l'élément mémoire OxRAM. On pourrait donc s'attendre aux mêmes conséquences, c'est à dire un risque d'une détérioration de l'uniformité de la distribution de la conductance en LRS due au manque de contrôle que l'on a sur celui-ci. Dans le cas présent, la capacité n'est plus simplement parasite, elle est ajoutée volontairement. On peut donc s'attendre à obtenir un meilleur contrôle de sa valeur qui devrait également être plus importante. L'amplitude du pic de courant n'est pas directement liée à la taille de la capacité[34], cependant, l'énergie contenue dans le pic l'est. On s'attendrait donc à constater un plus grand impact de ce pic dans le cas de l'écriture capacitive que dans celui utilisant une source de courant. Une limitation locale du courant pourrait donc s'avérer être nécessaire.

Cependant, la similarité avec l'écriture en courant s'arrête là. Dans le cas de la source de courant, avant le SET, la tension augmente, alors qu'elle diminue ici. Sachant que la probabilité de SET sur un intervalle de temps donné est proportionnelle à la tension sur l'élément mémoire OxRAM élevé à une certaine puissance [16], on s'attend à ce que la vitesse d'écriture puisse être maximale avec une source capacitive. Mais en contre-partie, une écriture à plus forte tension tend à causer plus de sur-écriture d'après les observations effectuées sur l'écriture avec une source de tension [51].

V.2.2 Contrainte d'intégration

Avec cette méthode d'écriture capacitive, la valeur de la capacité doit être suffisamment grande pour maintenir la tension assez haute assez longtemps pour que l'élément mémoire OxRAM puisse s'écrire correctement. Seulement, intégrer une capacité, voire plusieurs capacités, pour pouvoir faire des écritures en parallèle, peut rapidement s'avérer avoir un coût rédhibitoire en terme de surface. L'étude de Paola Trotti [79] couvre également la question de la taille de capacité requise. Elle conclut qu'une capacité de 330 pF chargée à une tension d'au moins 3 V est nécessaire. Bien qu'une capacité de cette taille soit tout à fait intégrable, elle consomme une surface équivalente à celle de la matrice de 16k éléments mémoires OxRAM dans la technologie du circuit *CAPA-WR*

[34]La hauteur maximale du pic de courant est égale au produit entre la tension à laquelle se produit le changement de conductance et la différence de conductance. Ce maximum n'est atteint que lorsque la dynamique du changement de conductance est bien supérieure à la dynamique de relaxation du couple RC. La capacité n'influence alors que l'énergie contenue dans ce pic en modulant le temps de relaxation de ce courant. Lorsque les dynamiques sont proches, le pic est « amorti » par la relaxation de couple RC avant d'atteindre son maximum [82].

Néanmoins, l'étude [79] utilise une capacité discrète (et donc externe au silicium contenant la matrice mémoire) connectée à l'élément mémoire OxRAM au travers de câbles coaxiaux et d'un testeur sous pointes. Il y a donc une forte probabilité que le partage de charges[35] entre la capacité discrète et l'ensemble des parasites diminue fortement la tension effective de départ ainsi que l'énergie disponible. De plus, comme seule l'énergie stockée à une tension supérieure à la tension du seuil d'écriture (environ 0,7 V sur l'échelle de temps de la microseconde) de l'élément mémoire OxRAM est utile, cela accentue d'autant plus la réduction de l'énergie utile disponible. Ainsi, nous pouvons nous attendre à ce que la capacité intégrée nécessaire soit plus faible. Nous retenons tout de même ce chiffre de 330 pF comme la quantité minimale de capacité potentiellement nécessaire.

V.2.3 Méthode pour la réduction de la capacité nécessaire

La forte contrainte sur la taille de la capacité provient de la nécessité qu'elle contienne l'intégralité de l'énergie de l'écriture. Cependant, on pourrait imaginer une stratégie pour relaxer cette contrainte.

On pourrait par exemple répéter l'opération d'écriture jusqu'à ce qu'elle ait lieu. Pour ce faire, on peut recharger la capacité dès que sa tension est trop faible. Cependant, l'élément mémoire OxRAM devrait être isolé de la capacité lors de sa recharge. Pour éviter cela, on peut utiliser une seconde capacité. En alternant la capacité que l'on recharge et la capacité que l'on utilise comme source sur l'élément mémoire OxRAM à chaque fois, on peut conserver en permanence une source sur l'élément mémoire OxRAM.

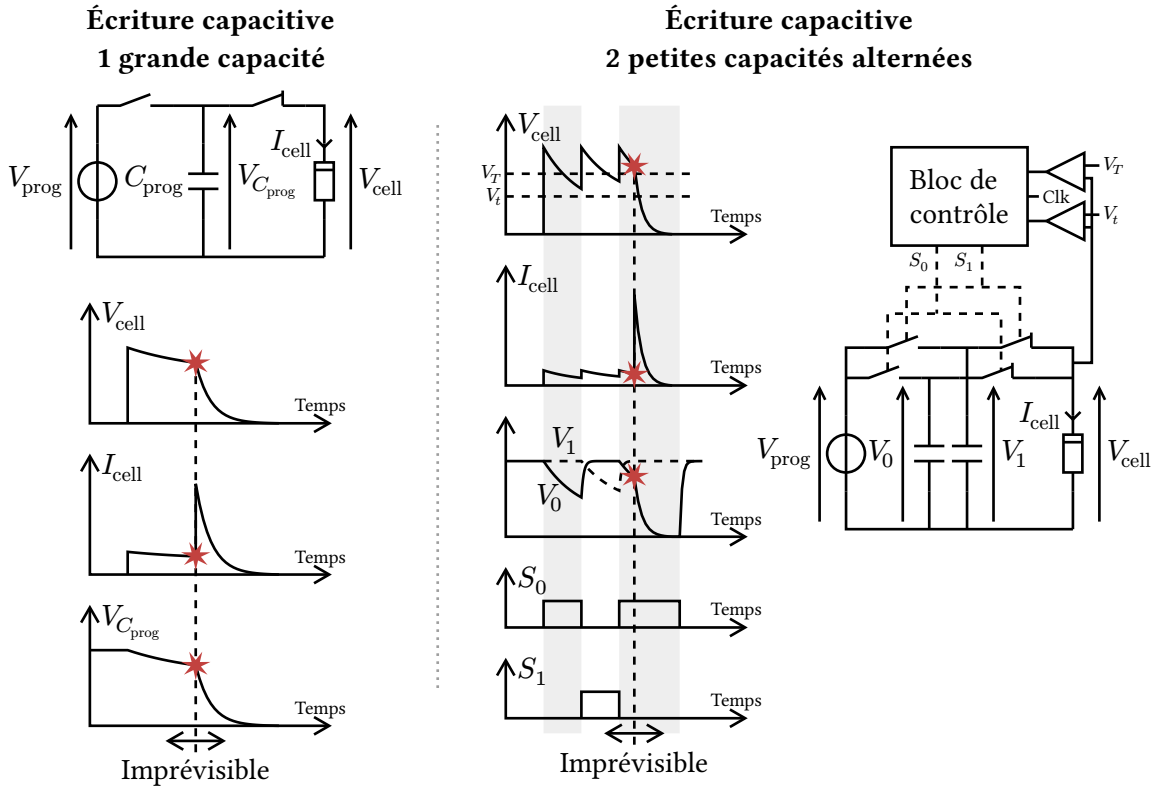


Figure 5.2 – Illustration de l'évolution de la tension sur l'élément mémoire OxRAM et du courant dans l'élément mémoire OxRAM pour une écriture avec une source capacitive statique et pour une écriture avec une source capacitive rechargée.

[35] Le phénomène de partage de charges a lieu lorsque deux éléments capacitifs sont connectés en parallèle alors que leur tension diffère. Lors de la connexion, la tension à leur borne tend à s'équilibrer rapidement. Pour ce faire, l'élément capacitif chargé à la tension la plus haute cède des charges à l'autre jusqu'à l'équilibre. La tension d'équilibre est la moyenne des tensions initiales pondérée par la capacité de chaque élément. On peut noter que de l'énergie est perdue lors de l'équilibrage.

Comme l'élément mémoire OxRAM n'a presque aucune chance de SET lorsque la tension est trop faible, on définit un seuil en-dessous duquel on échange les deux capacités. Du point de vue de l'élément mémoire OxRAM, cela a pour effet de recharger « instantanément » la capacité. Cependant, le principal intérêt de l'utilisation d'une capacité pour l'écriture est alors perdu. En effet, même lorsque le SET aura lieu, un tel système continuera à échanger les capacités, déversant toujours plus d'énergie sur l'élément mémoire OxRAM. La propriété d'auto-terminaison de l'écriture est perdue.

Pour remédier à cela, il est nécessaire de détecter le moment où l'élément mémoire OxRAM SET, pour arrêter le processus. Comme la résistance de l'élément mémoire OxRAM en LRS est beaucoup plus faible qu'en HRS, on peut par exemple effectuer cette détection en mesurant la vitesse de la décharge. La Figure 5.2 illustre une implémentation simpliste de cette idée qui est implémentée dans le circuit *CAPA-WR* discuté dans la prochaine partie. Deux seuils sont définis, un dit « haut » V_T et un autre dit « bas » V_t. Un signal d'horloge d'une période T est utilisé pour définir une base de temps. Lorsque durant deux coups d'horloge successifs, la tension est observée entre les deux seuils (à l'aide de comparateurs), les capacités sont échangées. Si la tension passe sous le seuil « bas », plus aucune recharge ne sera effectuée. Cela correspond à la situation où l'écriture serait finie. Avec cette méthode, on a la certitude que le système échangera les capacités tant que la résistance de l'élément mémoire OxRAM sera supérieure à $2R_0$ et que les capacités ne seront pas échangées dès qu'elle sera inférieure à R_0 (avec $R_0 = \frac{T}{2C}\frac{V_T+V_t}{V_T-V_t}$ et C la capacité d'écriture). Entre ces deux bornes, la réaction du système dépend de l'alignement entre la phase de l'horloge avec la décharge. Si l'on considère cette « phase » aléatoire et sans biais, la probabilité d'échanger les deux capacités lorsque la résistance de l'élément mémoire OxRAM R est entre R_0 et $2R_0$ est $\frac{R}{R_0} - 1$.

Cette implémentation n'est pas idéale. La plage d'incertitude est assez large. Cependant, il est important de trouver un bon équilibre entre précision et complexité. En effet, le principal avantage apparent de l'écriture capacitive est son auto-terminaison ne nécessitant pas de système supplémentaire. Avec cette méthode, nous avons en partie déjà perdu cet avantage. En effet, le système de détection du SET ressemble sur le principe aux systèmes de « terminaison d'écriture » utilisés pour l'écriture avec une source de tension. Pour garder un intérêt, notre système de détection doit rester simple et compact. L'implémentation proposée a le mérite de nous permettre d'expérimenter cette variation de la méthode d'écriture. L'intérêt de celle-ci pourra être déterminé plus tard, lorsque l'expérimentation aura éclairci l'intérêt ou les compromis liés à l'écriture capacitive.

V.3 Mise en œuvre de l'écriture capacitive : Circuit *CAPA-WR*

V.3.1 Objectifs

Comment dans le cas de l'étude de l'écriture en courant discutée au Chapitre IV, l'étude de l'écriture en charge nécessite la réalisation d'un circuit intégré dédié. En effet, la simulation est difficilement envisageable car il n'existe pas de modèle prédictif de l'écriture des éléments mémoires OxRAM. De plus, comme l'écriture en courant, l'écriture en charge est sensible à son environnement capacitif, c'est-à-dire, à la matrice de mémoire utilisée et aux interconnexions entre l'élément mémoire OxRAM et le circuit de programmation. Cela requiert donc une co-intégration de la source capacitive à une matrice d'éléments mémoires OxRAM, de préférence de taille suffisante pour pouvoir observer les effets parasites liés au routage interne à la matrice. La réalisation d'un circuit dédié permet de se rapprocher de la situation dans laquelle l'élément mémoire OxRAM serait utilisé en pratique.

Les contraintes sont assez similaires à celles du circuit *MEMORIIC* décrit dans le Chapitre IV.2.1 mais tout de même détaillées ci-dessous.

Le comportement des éléments mémoires OxRAM est très stochastique. Pour obtenir des données statistiquement représentatives, nous avons besoin d'un grand nombre d'éléments mémoires OxRAM.

L'utilisation d'une matrice de 16k éléments mémoires OxRAM nous permet d'avoir accès à un nombre suffisant d'élément mémoire OxRAM, pour une statistique à 4 σ.

Nous savons que la valeur de la capacité est un paramètre important pour l'écriture en charge. Nous avons donc besoin de la faire varier. Comme discuté plus tôt, il est nécessaire d'avoir cette capacité intégrée. Nous ne pouvons donc pas la faire varier en utilisant des composants externes comme cela était fait dans [79]. Nous avons besoin d'une banque de capacités intégrées dont on peut sélectionner la partie qui sera utilisée lors de l'écriture. Il existe également plusieurs technologies de capacité intégrée, avec souvent un compromis entre linéarité et densité. Il serait donc intéressant de pouvoir faire l'expérience avec ces différentes technologies. On souhaite également pouvoir mesurer les capacités ainsi intégrées. En effet, la fabrication n'est pas parfaitement homogène et la capacité effective peut varier entre deux puces et surtout entre deux galettes de silicium.

Pour permettre l'implémentation de la technique de recharge des capacités, il est nécessaire d'avoir deux exemplaires de la capacité d'écriture. La banque de capacités devra donc gérer cela. De plus, comme la matrice d'éléments mémoires OxRAM est un circuit dense, la capacité présente entre ses fils peut être non négligeable. Nous souhaitons donc pouvoir l'utiliser comme capacité d'écriture en plus de la banque de capacités intégrées. Enfin, la source de tension servant à charger la capacité d'écriture n'a quant à elle pas besoin d'être intégrée. Seule la tension de charge de la capacité devrait affecter l'écriture. Or, les parasites dus à la non-intégration de la source de tension n'affectent pas la tension de charge si l'on prend suffisamment de temps pour stabiliser la tension. Ne pas intégrer la source de tension simplifie également la conception et diminue les risques d'avoir un dysfonctionnement fatal empêchant l'utilisation du circuit. Nous préférons donc assurer le fonctionnement du circuit quitte à diminuer un peu la fréquence à laquelle nous effectuons les écritures.

Il est nécessaire de pouvoir comparer l'écriture en charge avec les autres techniques d'écriture. En particulier, nous souhaitons effectuer la comparaison avec l'écriture en tension. Il est préférable de réaliser cette comparaison sur le même circuit. Cela permet d'éliminer de nombreux biais comme la variabilité des processus de fabrication ou les différences d'environnement des éléments mémoires OxRAM. Il est donc nécessaire de prévoir un mode de fonctionnement pour l'écriture en tension. La source de tension pourra être externe ainsi que la commande numérique permettant de choisir la durée de l'écriture. Ce mode devra également pouvoir être utilisé pour effectuer la lecture de la conductance des éléments mémoires OxRAM de façon analogique (c'est-à-dire par mesure du courant au travers d'une faible source de tension appliquée aux bornes des éléments mémoires OxRAM). Encore une fois, dans notre contexte expérimental, aucune contrainte de vitesse n'est imposée tant que l'on peut effectuer les mesures dans un temps raisonnable. Ainsi, l'intégration de ce dispositif de lecture n'est pas nécessaire.

Le cadre dans lequel a été réalisé ce circuit imposait également des contraintes supplémentaires. Nous ne pouvons utiliser que 25 plots de connections externes et la largeur du circuit est limitée à celle des plots (2 mm). Ces limites sont imposées par le fonctionnement de la fabrication multi-projets qui utilise des barrettes de plots de largeur normalisée pour faciliter l'assemblage. Cela permet également l'utilisation d'un équipement de test sous pointes commun (sans requérir des cartes à pointes personnalisées et onéreuses). Nous limitons également la hauteur du circuit à la hauteur minimale (0,5 mm) requise pour insérer la barrette de plots ainsi que la matrice d'élément mémoire OxRAM pré-existante. La surface ainsi disponible est suffisante pour accueillir plus de capacité que nécessaire. En effet, au moins 0,5 mm² de surface est disponible pour intégrer des capacités. Dans la technologie utilisée, deux types de capacité peuvent être superposés, l'une avec une densité d'environ 3,4 nF/mm², l'autre avec une densité d'environ 0,5 nF/mm², donc pas loin de 2 nF devrait être disponible. Par contre, la limite sur le nombre de plots nécessite le multiplexage de nombreux signaux numériques de paramétrage.

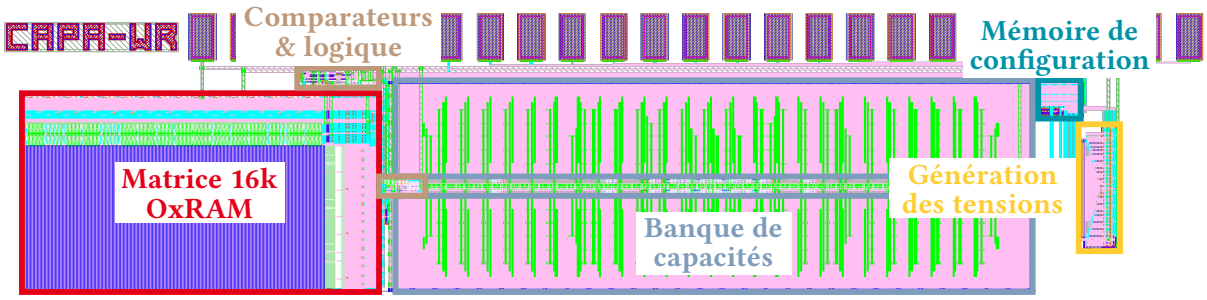


Figure 5.3 – Dessin des masques du circuit *CAPA-WR*.

Ainsi les objectifs du circuit *CAPA-WR* peuvent être résumés en :

1. Intégrer une matrice d'éléments mémoires OxRAM pour permettre l'obtention de résultats statistiquement représentatifs
2. Y co-intégrer des banques de capacités de différentes technologies permettant leur écriture en charge (possiblement avec rechargement)
3. Permettre la mesure de ces capacités intégrées
4. Permettre l'écriture des éléments mémoires OxRAM par une source de tension externe
5. Permettre une lecture analogique de la conductance de chaque élément mémoire OxRAM
6. Ne pas utiliser plus de 25 connexions externes (alimentation incluse)

V.3.2 Présentation du circuit *CAPA-WR*

La Figure 5.3 montre le dessin final du circuit *CAPA-WR* réalisé dans une technologie CMOS 130 nm industrielle hybridée avec la technologie OxRAM du CEA-Leti. Ce circuit est composé :

- D'une matrice 16k d'éléments mémoires OxRAM, la même que celle utilisée dans le circuit *MEMORIIC*
- De deux banques de capacités identiques entre elles contenant des capacités MOM et MOS
- De deux générateurs de tension de seuil ainsi que deux comparateurs pour permettre la mise en œuvre de la technique de recharge des capacités
- D'une mémoire de configuration
- De la logique permettant le contrôle du circuit

De nombreuses portes de transmission sont utilisées dans le circuit pour permettre la reconfiguration des connexions internes (voir Annexe B.3). Cette logique de reconfiguration a été conçue pour supporter les différents modes de fonctionnement identifiés :

1. Ecriture des éléments mémoires OxRAM avec une source de tension externe (Annexe B.4)
2. Lecture analogique du niveau de conductance des éléments mémoires OxRAM (Annexe B.5)
3. Mesures des capacités intégrées (Annexe B.6)
4. Ecriture en charge des éléments mémoires OxRAM (Annexe B.7)
5. Ecriture en charge des éléments mémoires OxRAM avec rechargement (Annexe B.8)

Elle permet également des variations telles que :

- Conserver la connexion externe vers la capacité intégrée lors de l'écriture pour permettre une éventuelle mesure de l'évolution temporelle du signal ou l'ajout d'une capacité externe
- Précharger les lignes de la matrice avant l'écriture pour éviter le phénomène de partage de charge
- Utiliser la capacité parasite des lignes de la matrice qui ne sont pas utilisées pendant l'écriture en tant que capacité d'écriture
- Utiliser les deux banques de capacités à la fois pour doubler la capacité disponible

Toutes ces options doivent correspondre à des signaux permettant à l'utilisateur de les sélectionner ou non. Cela génère donc un grand nombre de signaux numériques de configuration. Au total, 50 bits de

configuration sont nécessaires. Pour ne pas dépasser l'objectif de 25 plots de connections externes, le choix de multiplexer ces signaux a été fait. En effet, ils sont à la fois nombreux et simples à multiplexer car ils n'ont pas de contrainte de timing et sont simplement mémorisables. Ce sont donc de bons candidats pour commencer. En effectuant le compte des signaux analogiques : 2 pour l'alimentation, 3 pour la connexion à une cellule 1T1R et 1 pour fixer le potentiel des cellules non connectées, on constate qu'il reste un budget de 19 signaux numériques. Le signal définissant le temps d'écriture ainsi que celui de l'horloge (nécessaire à la technique de recharge des capacités) doivent avoir un timing précis. On leur réserve donc un plot chacun. Il reste 17 plots pour les 50 bits de configuration. On a donc assez de marge de manœuvre pour implémenter une interface parallèle constituée d'un bus de données et d'un bus d'adresses pour communiquer avec la mémoire interne. Comme l'adressage de la matrice se fait sur 14 bits, il est confortable d'utiliser un bus de données de 14 bits et 3 bits sont encore disponibles pour l'adressage ou le contrôle. Comme 4 registres de 14 bits (56 bits) suffisent, 2 bits sont utilisés pour l'adressage et 1 pour activer ou désactiver l'écriture du registre adressé. L'assignation finale des 25 plots est donnée en Annexe B.1 et le fonctionnement détaillé de la mémoire de configuration ainsi que la description des registres sont donnés en Annexe B.2.

Les capacités intégrées sont séparées en deux banques identiques. Chaque banque contient à la fois des capacités MOM et des capacités MOS. Les capacités MOS sont des transistors MOSFET dont la source, le drain et le substrat forment un pôle de la capacité connectée à la masse et dont la grille forme le second pôle. Cette technologie présente une forte densité capacitive mais sa caractéristique $C(V)$ n'est pas constante. De façon simplifiée, la capacité de l'oxyde entre la grille et le substrat n'est effective que lorsque la couche d'inversion est formée, c'est-à-dire lorsque la tension est plus haute que la tension de seuil du transistor. Lorsque la tension n'est pas assez forte, la capacité effective n'est qu'une fraction de la capacité de l'oxyde de grille car elle n'est électriquement accessible qu'au travers de la capacité due à la zone de déplétion. Dans notre cas, cette non-linéarité pourrait être bénéfique puisque qu'elle concentre l'énergie vers les tensions supérieures à la tension de seuil. Une fois la couche d'inversion formée, la capacité surfacique des capacité MOS dans cette technologie est d'environ 3,4 nF/mm².

Les capacités MOM sont constituées de deux armatures métalliques inter-digitées réalisées à l'aide des niveaux métalliques normalement utilisés pour le routage. Leur caractéristique est plutôt linéaire mais leur densité, environ 0,5 nF/mm², est plus faible. Comme les capacités MOS n'utilisent que les niveaux substrat, poly ainsi que le premier niveau de métal, les niveaux de métal deux, trois et quatre peuvent être utilisés pour faire des capacités MOM directement au-dessus des capacités MOS. Ainsi, chaque banque de capacités intègre une surface équivalente de capacités MOM et MOS.

14 bits de configuration sont réservés à la sélection de la valeur de capacité souhaitée par l'utilisateur. Ces 14 bits sont répartis en deux groupes de 7 bits correspondants aux deux types de technologie. Cela permet de sélectionner uniquement l'une ou l'autre technologie mais aussi de mélanger les deux. Il y a donc 128 valeurs de capacité disponible pour chaque technologie. Le dimensionnement des capacités a été effectué « à l'envers », c'est-à-dire en partant de la surface disponible et de la granularité souhaitées. Cela nous a conduit à une valeur unitaire de 0,6 pF MOM et 3,8 pF MOS, pour un total de 76 pF MOM et 480 pF MOS disponible (ou le double si les deux banques sont utilisées en même temps). Cela dépasse donc la valeur minimum de 330 pF issue de [79] que l'on s'était fixée comme objectif.

Le circuit *CAPA-WR* a été conçu avec beaucoup d'inconnues et de supposition. Il a donc été conçu pour être versatile. Cela permettra, on l'espère, de répondre au mieux à tous les besoins expérimentaux qui se révéleront au fur et à mesure de son exploitation.

V.3.3 Simulation

Le circuit *CAPA-WR* offre différents modes de fonctionnement. Pour s'assurer que l'ensemble des modes soit fonctionnel, un ensemble de simulations a été mené. Ces simulations, communément

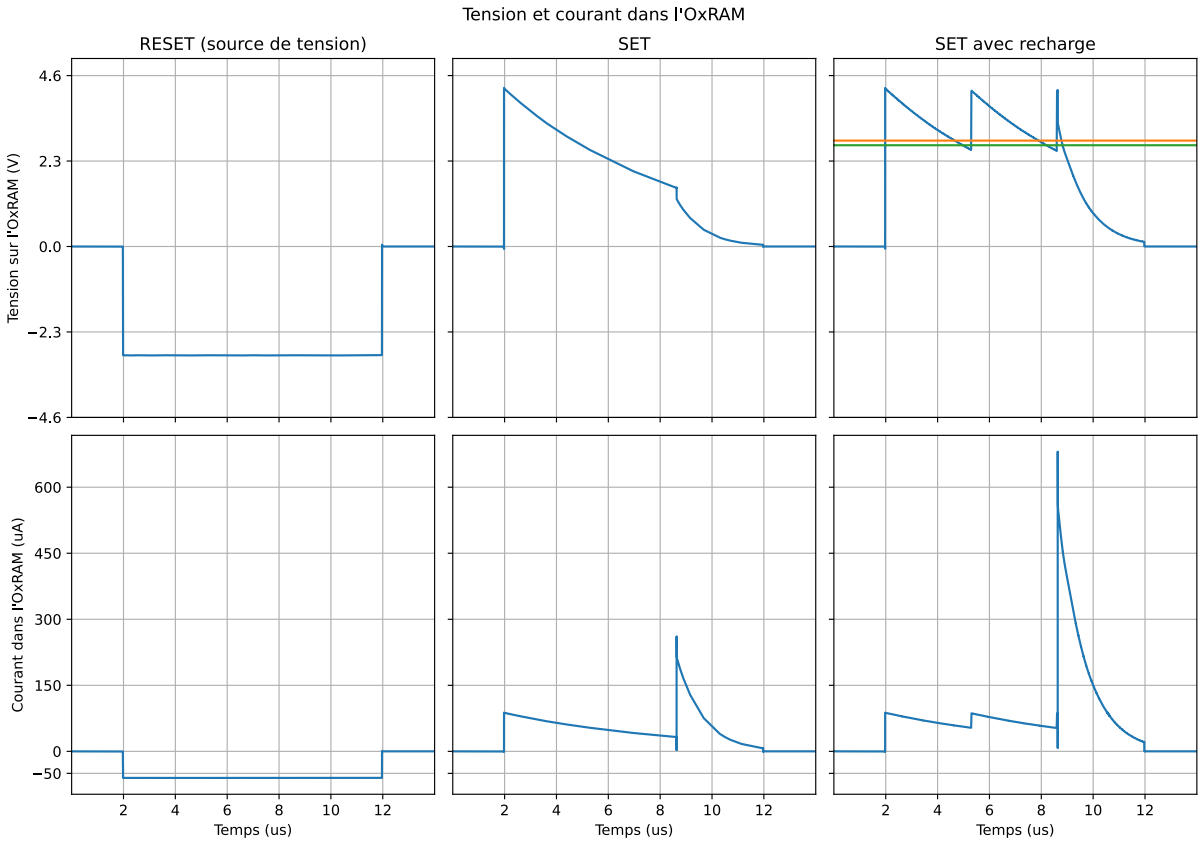


Figure 5.4 – Tension et courant dans l'élément mémoire OxRAM dans 3 simulations d'écriture. De gauche à droite, avec une source de tension et en RESET, avec une source capacitive, avec une source capacitive et le système de rechargement de la capacité. Dans ces simulations, l'élément mémoire OxRAM est représenté par un élément résistive de 50 kΩ en HRS changeant brusquement de résistance à un instant prédéterminé pour atteindre 5 kΩ en LRS. La capacité utilisée est composée de 120 x 0,6 pF MOM et 5 x 3,8 pF MOS pour un total de 91 pF.

qualifiées de simulation TOP, incluent l'ensemble du circuit ainsi qu'un modèle du banc de test. Elles ont pour intérêt de tester l'ensemble du circuit pour vérifier que la logique corresponde bien à ce que l'on avait prévu et décrit dans la documentation. Dans ces simulations, l'élément mémoire OxRAM est remplacé par une résistance dont la valeur évolue pendant la simulation de façon prédéterminée. Ainsi, elles ne permettent pas de vérifier l'écriture de l'élément mémoire OxRAM mais uniquement le fonctionnement du circuit. Aussi, la présence de la matrice 16k ralentit fortement les simulations, particulièrement lorsque l'horloge nécessaire à l'écriture avec recharge est active. Lorsque cela est nécessaire, la netlist de la simulation a été simplifiée en retirant les lignes et les colonnes inutiles de la matrice, à l'exception de la ligne et de la colonne sélectionnées ainsi qu'une de chaque non sélectionnée pour pouvoir vérifier que les dispositifs non sélectionnés ne sont pas écrits par erreur.

La Figure 5.4 illustre le résultat de trois simulations. La tension sur l'élément mémoire OxRAM ainsi que le courant passant dedans sont représentés. La première, à gauche, vérifie notre capacité à utiliser une source de tension pour l'écriture (ici pour le RESET). La seconde montre une écriture capacitive sans rechargement et la dernière illustre l'écriture capacitive avec rechargement. On peut constater que la hauteur du pic de courant est bien proportionnelle à la tension à laquelle le SET se produit.

L'ensemble des simulations nous a permis de vérifier le fonctionnement du circuit dans toutes les configurations prévues pour le test et l'exploitation du circuit.

V.4 Conclusion

Le circuit *CAPA-WR* n'a pu être testé ni exploité pendant cette thèse, puisque sa réception était prévue pour la fin de l'année 2024, au même moment que la fin de cette thèse. Cependant, les vérifications par simulation de l'ensemble du circuit pour les différents modes d'utilisation envisagés permettent d'être confiant sur son bon fonctionnement. Mais, sans l'exploitation de ce circuit, aucune conclusion ne peut être donnée sur l'écriture capacitive.

D'un point de vue théorique, nous avons expliqué pourquoi l'écriture capacitive est intéressante pour résoudre le problème de terminaison de l'écriture et de la sur-écriture associée. Nous avons également discuté des contraintes d'intégration liées à la capacité et d'une méthode permettant de réduire la taille de la capacité en échange d'une augmentation de la complexité du système. Pour que l'écriture capacitive soit considérée intéressante, il faudrait que l'expérience montre suffisamment d'intérêt pour contre-balancer ces contraintes ou prouve qu'une faible capacité est suffisante.

Dans un contexte applicatif, la surface d'un module de programmation est un facteur critique pour l'attractivité d'une méthode. En effet, il est nécessaire d'intégrer plusieurs modules pour permettre l'écriture de plusieurs mémoires simultanément. La surface du module peut donc rapidement représenter une part importante de la surface globale. Cependant, sans les résultats de l'expérimentation, et donc sans la taille de la capacité nécessaire connue, il est difficile de tirer une conclusion. La surface nécessaire sera un facteur clé de l'attractivité ou non de l'écriture capacitive.

L'aspect sur lequel cette méthode devrait se différencier est la faible quantité d'énergie utilisée pour écrire un bit. Cependant, il faudra ne pas oublier de prendre en compte la charge de la capacité qui peut être source d'une perte d'énergie. Un système de recharge de capacité tendant vers l'adiabatique pourrait être utilisé pour améliorer l'efficacité.

Bibliographie

[16] G. Sassine, C. Cagli, J.-F. Nodin, G. Molas, et E. Nowak, « Novel Computing Method for Short Programming Time and Low Energy Consumption in HfO2 Based RRAM Arrays », *IEEE Journal of the Electron Devices Society*, vol. 6, p. 696-702, 2018, doi: 10.1109/JEDS.2018.2830999.

[25] W.-H. Chen *et al.*, « A 16 Mb dual-mode ReRAM macro with sub-14 ns computing-in-memory and memory functions enabled by self-write termination scheme », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 28. doi: 10.1109/IEDM.2017.8268468.

[51] Z.-Q. Wang *et al.*, « Cycling-induced degradation of metal-oxide resistive switching memory (RRAM) », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 7. doi: 10.1109/IEDM.2015.7409649.

[79] P. Trotti, G. Pillonet, G. Molas, S. Oukassi, et E. Nowak, « Experimental Set-Up For Novel Energy Efficient Charge-based Resistive RAM (RRAM) Switching », in *2 020 IEEE 33rd International Conference on Microelectronic Test Structures (ICMTS)*, mai 2020, p. 1-5. doi: 10.1109/ICMTS48187.2020.9107936.

[82] P. R. Shrestha *et al.*, « Analysis and Control of RRAM Overshoot Current », *IEEE Transactions on Electron Devices*, vol. 65, nº 1, p. 108-114, janv. 2018, doi: 10.1109/TED.2017.2776860.

Chapitre VI : Limites de l'augmentation de la densité

VI.1 Introduction

Les précédents chapitres constituent une partie concentrée sur l'écriture des éléments mémoires OxRAM. Ce chapitre et le suivant traitent de la question de la densité.

La densité est un critère très important pour les mémoires. C'est, par exemple, le principal facteur modulant le coût de la mémoire. Pour qu'une mémoire soit compétitive, ou ait l'espoir de le devenir, elle doit pouvoir proposer une densité importante, ou du moins avoir une perspective d'amélioration de sa densité. Bien évidemment, la densité n'est pas le seul critère. Par exemple, la vitesse d'écriture et la capacité à être co-intégrée sur une technologie CMOS représentent un avantage non négligeable pour l'OxRAM par rapport à une mémoire Flash telle qu'utilisée dans les SSD. Ainsi, il n'est pas question de rivaliser avec leur densité. De l'autre côté du spectre, la SRAM a pour principal défaut d'être limitée en densité, en contrepartie, elle est particulièrement rapide. L'OxRAM doit donc pouvoir proposer une densité supérieure pour être intéressante[36], sachant qu'elle sera moins rapide. La eFlash, pour embedded flash, présente des caractéristiques proches de ce que peut offrir l'OxRAM. Elle représente donc un point de comparaison intéressant. Son principal défaut est qu'elle nécessite une alimentation d'environ 10 V pour l'écriture. A performances équivalentes, l'OxRAM serait donc une meilleure alternative. Le Tableau 6.1 résume les caractéristiques principales de ces mémoires.

Tableau 6.1 – Tableau comparatif de différentes technologies mémoire

Mémoire	SRAM	eFlash [88]	DRAM	NAND Flash [89]	NOR Flash [89]
Densité (F^2)	~ 300[α]	–	6 – 8	4	10
Densité (μm^2)	~ 0.1[α]	0.053	–	–	–
Temps de lecture (us)	~ 0.005	0.05	0.1	25	0.075
Temps d'écriture (us)	~ 0.005	30	0.1	220 par 2 112 bytes	128 par 32 bytes
Temps d'effacement (us)				500 par 128kW	1 000 par 128kW
Rétention	Volatile	Persistant	Volatile	Persistant	Persistant
Intégrable CMOS	Oui	Oui	Non	Non	Non

[α] Valeurs pour un nœud CMOS de 22 nm. Pour les nœuds plus avancés, la densité en F^2 ne peut être maintenue et augmente à 600 F^2 pour un nœud 7 nm par exemple [90].

[36] L'OxRAM a également l'avantage d'être non volatile, ce qui peut être un prérequis dans certaines applications. Dans ce cas, elle n'est pas en compétition avec la SRAM ou la DRAM.

La densité de l'OxRAM est fortement liée à son intégration. Une grande densité requiert une faible surface par cellule mémoire. C'est pourquoi nous nous intéresserons d'abord aux tendances d'évolution des propriétés de l'élément mémoire OxRAM lorsque sa surface est réduite. La nécessité d'intégrer un grand nombre d'éléments mémoires OxRAM sur une surface limitée pose un problème logistique, comment accéder électriquement à chaque élément ? C'est cette question qui sera au cœur de cette partie.

Nous commençons par nous interroger sur ce que signifie « accès électrique » plus en détail. Nous identifions ensuite deux grands types de cellule mémoire et discutons de la mise en œuvre de leur intégration pour permettre l'accès à chacune.

VI.2 Dispositif unitaire

VI.2.1 Intégration

L'élément mémoire OxRAM est intégré dans le *back-end-of-the-line* du processus de fabrication CMOS, c'est-à-dire dans la partie du processus qui forme et définit les pistes métalliques qui connectent électriquement les dispositifs silicium formés sur le substrat lors de la partie *front-end-of-the-line* du processus. Les éléments mémoires OxRAM sont fabriqués entre deux niveaux de métal comme un via métallique. Cela limite donc le budget thermique disponible à la fabrication de l'élément mémoire OxRAM à celui encore disponible à ce moment du processus.

VI.2.2 Lien entre surface, temps de formation et tension de formation

Nous nous concentrons dans un premier temps sur l'élément mémoire OxRAM seul, sans prendre le sélecteur ou autre périphérie.

Dans l'état le plus conducteur de l'élément mémoire OxRAM, le LRS, nous avons vu au Chapitre II que la conduction est dominée par celle dans le filament formé par les lacunes d'oxygène. Ce mécanisme ne met en jeu qu'une faible partie de la surface de l'élément mémoire OxRAM. Ainsi, l'état de conductance atteint en LRS est indépendant de la surface de l'élément. En HRS, le filament n'est généralement pas entièrement dissous. Le mécanisme de conduction dominant est alors la conduction au travers de l'oxyde du bout du filament vers l'électrode. Ainsi, en HRS également, la caractéristique de conductance de l'élément mémoire OxRAM est assez indépendante de la surface de l'élément.

Cependant, la surface joue sur certains aspects du fonctionnement de l'élément mémoire OxRAM. L'état de conductance à l'état neuf, c'est-à-dire lorsqu'aucun filament n'a encore été formé, la conductance de l'élément est proportionnellement reliée à sa surface, comme dans n'importe quelle structure métal-oxyde-métal.

Lors du Forming, l'ensemble de la surface est initialement mis à contribution. Pour que le Forming se produise, il est nécessaire de générer une accumulation de lacunes d'oxygène jusqu'à former un filament. Comme nous l'avions vu au Chapitre II, la vitesse du processus de formation des lacunes est amplifiée par la présence d'autres lacunes. C'est d'ailleurs ce qui expliquait la formation d'un filament localisé, et non pas d'un nuage de lacunes uniformément distribuées. Ce processus est similaire à la rupture de l'oxyde étudié pour des questions de fiabilité de MOSFET.

Pour établir la relation entre le temps de rupture/formation, considérons n éléments mémoires OxRAM connectés électriquement en parallèle. Ils subissent alors tous la même tension. Dès que l'un de ces éléments se forme, l'ensemble connecté en parallèle apparaît comme formé pour l'opérateur. Ainsi, le temps de formation T de l'ensemble est le minimum des temps de formation des éléments le constituant T_i. On suppose les T_i identiquement et indépendamment distribués selon une loi de Weibull[37]

[37]Le choix de cette distribution qui est l'une des trois lois d'extremum généralisée, particulièrement celle de support $\mathbb{R}^+$ se justifie par le fait que nous étudions un problème de rupture. Si l'on considère que le temps d'apparition des événements à

de paramètre de forme k et paramètre d'échelle λ (comme fait pour le temps de rupture d'un oxyde [91]). T suit alors également une loi de Weibull de même paramètre de forme, mais de paramètre d'échelle $n^{-\frac{1}{k}}\lambda$. L'espérance de T est donc $E(T) = n^{-\frac{1}{k}}E(T_i)$. Donc l'ensemble des n éléments, qui a une surface cumulée n fois plus grande qu'un élément seul, a un temps de formation moyen divisé par $n^{\frac{1}{k}}$. En considérant que nos n éléments mémoires OxRAM en parallèle sont équivalents à un unique élément ayant une surface n fois plus grande, on conclut que plus la surface d'un élément mémoire OxRAM est grande, plus sa formation est rapide, à tension fixe bien sûr.

La conséquence est que la réduction de la surface des éléments mémoires OxRAM, pour permettre leur intégration plus dense, accroît le temps de formation. Ce temps peut être récupéré en augmentant la tension, ce qui accélère le processus de formation [60]. Cependant, la tension disponible est un facteur limitant, d'autant plus qu'une intégration plus dense requiert un nœud CMOS plus avancé et donc supportant des tensions maximales plus faibles.

Pour essayer de compenser cette augmentation de la tension de formation, des modifications de la composition du matériau de l'oxyde ont été essayées. L'implantation de silicium dans l'oxyde d'hafnium montre de bons résultats dans ce sens [76] [77]. L'utilisation de divers schémas d'impulsion électrique, en particulier visant à provoquer une rapide différence de potentiel d'une amplitude plus grande que la tension maximale admissible [92], donne également des résultats encourageants pour diminuer la tension de formation.

VI.3 Multiplexage des accès

L'élément mémoire OxRAM est un dipôle. On peut donc imposer la tension sur l'élément à l'aide de deux lignes électriques et d'une source. Cependant, pour réaliser une mémoire, un nombre important de dispositifs doit être accessible. Il n'est alors pas possible de conserver deux lignes, ou même une seule, par éléments. Au-delà du problème logistique, cela serait tout simplement un problème géométrique car on ne peut avoir qu'un certain nombre de lignes électriques par unité de surface. De plus, il faudra bien agir sur ces lignes, pour modifier leur potentiel par exemple. Pour cela il sera nécessaire d'avoir autant d'actionneurs que de lignes. La densité effective, prenant en compte l'élément mémoire OxRAM ainsi que sa périphérie, serait alors fortement diminuée. Il convient donc de trouver un moyen de multiplexer les accès.

Dans le circuit *WT-PAOLA*, les éléments mémoires OxRAM étaient accessibles au travers de multiplexeurs analogiques (portes de transmission). Cela reviendrait donc à avoir au moins quatre transistors par élément mémoire OxRAM (sans compter la logique d'adressage). Ici, la quantité de périphérie requise croit linéairement avec le nombre d'éléments mémoires OxRAM. Seuls quatre éléments mémoires OxRAM étaient présents dans le circuit *WT-PAOLA*, cette solution ne posait donc pas de problème. Cependant, pour assurer une grande densité effective lorsque le nombre de dispositifs est grand , il est nécessaire de trouver une solution présentant une croissance sous-linéaire. Pour cela, il faudrait le plus possible partager la périphérie. Cependant, il est nécessaire de conserver la capacité d'accéder à chaque élément indépendamment des autres, c'est ce que l'on appelle sélectionner un élément.

VI.3.1 Adressage d'un dipôle

Pour sélectionner un dipôle, il est nécessaire d'adresser ses deux pôles simultanément, et indépendamment des autres. Imaginons avoir N éléments mémoires OxRAM à connecter ensemble tout en pouvant les sélectionner individuellement.

l'échelle microscopique est identiquement et indépendamment distribué, peu importe leur distribution, celle de leur minimum qui nous intéresse ici, suit nécessairement une loi d'extremum généralisée si elle existe.

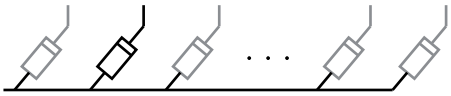


Figure 6.1 – Arrangement avec un des deux pôles conducteurs mis en commun entre toutes les cellules. Il y a $N + 1$ connections externes avec N le nombre d'éléments mémoires OxRAM.

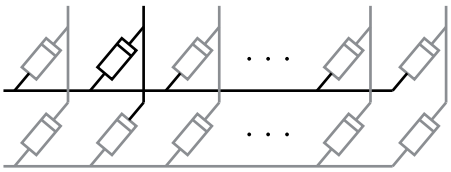


Figure 6.2 – Arrangement avec deux groupes au sein desquels l'un des deux pôles conducteurs est mis en commun, et où le pôle restant est connecté avec l'homologue de l'autre groupe. Il y a $\frac{N}{2} + 2$ connections externes avec N le nombre d'éléments mémoires OxRAM.

Comme illustré sur la Figure 6.1, on pourrait connecter ensemble l'un des pôles de chaque élément mémoire OxRAM, il est alors possible de réduire le nombre de ligne à $N + 1$ et de sélectionner chaque élément au travers de son pôle qui n'est pas connecté à la ligne commune.

On pourrait faire de même, mais en utilisant deux lignes communes différentes, en répartissant les éléments en deux groupes. C'est ce qu'illustre la Figure 6.2. Le pôle libre de chaque élément pourrait alors être connecté sur la ligne également connectée au pôle libre d'un élément de l'autre groupe. Cela demanderait $\frac{N}{2} + 2$ lignes. La sélection se ferait alors en utilisant la ligne commune correspondant au groupe de l'élément souhaité, ainsi que la ligne connectée à la paire contenant cet élément. Cette répartition réduit d'environ moitié le nombre de lignes nécessaires.

On peut continuer dans cette direction jusqu'à la limite où $\sqrt{N}$ groupes de $\sqrt{N}$ éléments sont considérés. On obtient alors un nombre de lignes égales à $2\sqrt{N}$. Cette division est optimale pour des dipôles que l'on souhaite pouvoir sélectionner, l'utilisation de moins de $2\sqrt{N}$ lignes ne permettrait pas de sélectionner les différents éléments mémoires OxRAM[38]. Le moyen le plus simple de réaliser de telles interconnexions est de positionner les éléments mémoires OxRAM de manière régulière sous forme

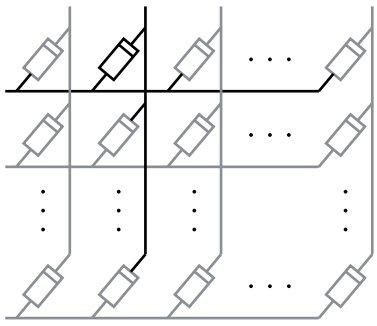


Figure 6.3 – Arrangement sous forme de matrice carrée. Les N cellules sont réparties sur un tableau de $\sqrt{N}$ lignes et $\sqrt{N}$ colonnes. Il y a $2\sqrt{N}$ connexions externes avec N le nombre d'éléments mémoires OxRAM.

[38]Soit un ensemble $\mathbb{E}$ de cardinal $N = n^2$ et $S(\mathbb{E})$ l'ensemble des singletons de $\mathbb{E}$, et deux partitions A et B de $\mathbb{E}$ de cardinal respectif n_A et n_B tel que $S(\mathbb{E}) \subset \{a \cap b \mid a \in A, b \in B\}$. $\mathbb{E}$ représente les éléments mémoires OxRAM à adresser, chaque ensemble dans A et B représente une ligne connectant les éléments qu'il contient. La condition d'inclusion de $S(\mathbb{E})$ représente la nécessité de pouvoir choisir deux lignes sélectionnant un élément mémoire OxRAM pour tous les éléments de $\mathbb{E}$. Par l'inclusion, $n^2 = |S(\mathbb{E})| \leq n_A n_B$. Donc $n_B \geq \frac{n^2}{n_A}$, en ajoutant n_A on obtient $n_A + n_B \geq \frac{n_A^2 + n^2}{n_A}$. En notant $f(x) =$

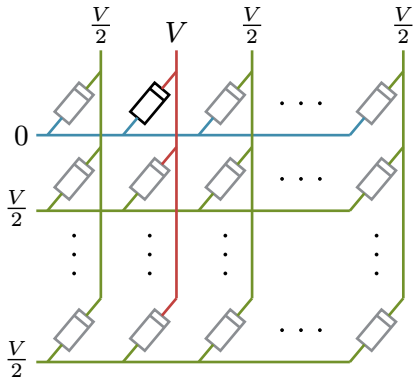


Figure 6.4 – Règle du 1/2 appliquée à la matrice. La cellule sélectionnée est la seule à voir une tension V. Les $2\sqrt{N}$ cellules partageant la ligne ou la colonne de cette dernière voient une tension de $\frac{V}{2}$. Les autres $\left(\sqrt{N}-1\right)^2$ cellules voient une tension nulle. Cette règle minimise le nombre de cellules non sélectionnées voyant une tension non nulle.

d'un tableau de $\sqrt{N}$ lignes par $\sqrt{N}$ colonnes, comme illustré sur la Figure 6.3. Puis de connecter l'un des pôles de chaque élément à celui des éléments de la même ligne, et l'autre pôle à celui des éléments de la même colonne. Cela constitue ce que l'on appelle communément une matrice de cellule mémoire, la cellule n'étant, ici, composée que de l'OxRAM.

Cette organisation est efficace et permet de sélectionner chaque élément mémoire OxRAM, mais il y a un petit problème. Imaginons que nous souhaitons appliquer une tension V sur la cellule de la ligne x et de la colonne y. Pour ce faire, on peut fixer la ligne x à la masse et la colonne y à la tension V. La question est maintenant de fixer le potentiel des autres lignes et colonnes[39]. Idéalement, il faudrait que la tension sur toutes les autres cellules soit nulle, mais c'est impossible. En effet, il serait nécessaire d'avoir, d'un côté, les autres lignes au même potentiel que la colonne y, c'est-à-dire V, et d'un autre côté, les autres colonnes à celui de la ligne x, c'est-à-dire $0V$. Cependant, cela induit que les cellules qui ne sont ni sur la ligne x, ni sur la colonne y verraient une tension $-V$. Le mieux que l'on puisse faire est d'essayer de répartir cette tension sur les différentes cellules non-sélectionnées pour qu'elles

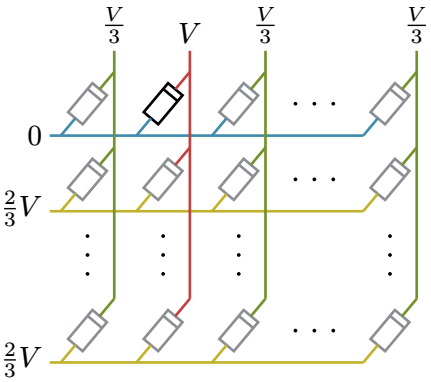


Figure 6.5 – Règle du 1/3 appliquée à la matrice. La cellule sélectionnée est la seule à voir une tension V. Les $2\sqrt{N}$ cellules partageant la ligne ou la colonne de cette dernière voient une tension de $\frac{V}{3}$. Les autres $\left(\sqrt{N}-1\right)^2$ cellules voient une tension de $-\frac{V}{3}$. Cette règle minimise la valeur absolue de la tension parasite apparaissant sur les cellules non sélectionnées.

$\frac{x^2+n^2}{x}$, on a $f'(x) = 1 - \frac{n^2}{x^2}$ qui s'annule en $x = \pm n$ et $f''(x) = 2\frac{n^2}{x^3} > 0$ pour $x > 0$, donc $f(x) > f(n) = 2n$. Donc $n_A + n_B \geq \frac{n_A^2+n^2}{n_A} \geq 2n$ qui est atteint pour $n_A = n_B = n$.

[39] Laisser les lignes flottantes est une autre option. Cependant, elle est équivalente (en statique) à avoir fixé chaque ligne à son potentiel d'équilibre.

voient une tension plus faible que celle sélectionnée. Pour cela il y a deux stratégies communément utilisées :

1. **Règle du 1/2** (voir Figure 6.4)
   On peut fixer chaque ligne et chaque colonne non sélectionnée à $\frac{V}{2}$. Cela induit une tension nulle sur les cellules ne partageant ni la ligne x, ni la colonne y et une tension $\frac{V}{2}$ sur les cellules présentes sur la ligne x ou sur la colonne y. Cette règle minimise le nombre de cellules non sélectionnées voyant une tension non nulle.
2. **Règle du 1/3** (voir Figure 6.5)
   On peut fixer chaque ligne restante à $\frac{2}{3}V$ et chaque colonne restante à $\frac{V}{3}$. Cela induit une tension $\pm\frac{V}{3}$ sur toutes les cellules hormis celle sélectionnée. Cette règle minimise la valeur absolue de la tension parasite apparaissant sur les cellules non sélectionnées.

Dans cette architecture, la capacité à sélectionner une unique cellule est donc toute relative. Cependant, comme la vitesse de réaction de l'élément mémoire OxRAM n'est pas liée de manière linéaire à la tension, il reste possible d'écrire un dispositif spécifique en conservant une faible probabilité d'écrire les autres cellules. Cette architecture est souvent dénommée « crossbar ».

Comme l'élément mémoire OxRAM est essentiellement une résistance variable, toute tension sur une cellule correspond à un courant la traversant. En HRS, le courant de la cellule dépend de la tension de façon exponentielle, ces éléments ne génèrent donc qu'un faible courant parasite. Cependant, en LRS, la conduction de l'élément mémoire OxRAM est approximativement ohmique, le courant est donc proportionnel à la tension. Ainsi, avec ce mode de fonctionnement, chaque élément en LRS génère un courant parasite du même ordre de grandeur que le courant utile. Ce courant représente, dès une taille de matrice de quelques cellules par quelques cellules, la majorité du courant injecté dans la matrice. Ainsi les tailles de matrices réalisables sont limitées par ce courant, et les circuits périphériques doivent être dimensionnés pour les fournir. La densité effective incluant les cellules et la périphérie en est d'autant pénalisée.

La lecture de la conductance d'une cellule est également perturbée par ces courants parasites.

Pour limiter ce courant parasite, il est nécessaire d'ajouter un dispositif dans chaque cellule, c'est ce que l'on appelle un sélecteur. Une telle cellule est qualifiée de cellule 1S1R.

VI.3.2 Cellules 1S1R

Dans une technologie CMOS, la surface active de silicium est précieuse. C'est pourquoi nous nous intéressons en premier lieu aux sélecteurs que l'on pourrait intégrer, tout comme l'élément mémoire OxRAM, entre deux niveaux de métal dans le back-end-of-the-line du processus de fabrication CMOS.

Comme la sélectivité de notre adressage se joue sur l'amplitude, c'est-à-dire que la cellule sélectionnée est la seule à voir une certaine tension, pendant que les autres voient au plus une fraction de cette tension, le sélecteur doit présenter une forte non-linéarité pour prévenir la conduction lorsque la tension est en-dessous d'un certain seuil, comme une diode, tout en permettant la circulation d'un courant suffisant lorsqu'il est passant. Comme l'élément mémoire OxRAM est bipolaire, il est nécessaire que le sélecteur puisse conduire du courant dans les deux sens. Cela élimine donc la diode.

Un point important à prendre en considération est la chute de tension sur le sélecteur. En effet, même si le sélecteur présente une non-linéarité franche permettant d'isoler les cellules à moitié sélectionnées, il n'est pas nécessairement utilisable. Il faut vérifier qu'une fois rendu passant la tension à ses bornes soit suffisamment faible pour que la tension restante, qui tombe donc sur l'élément mémoire OxRAM, soit suffisante pour écrire la mémoire. Si ce n'est pas le cas, il peut être nécessaire d'augmenter la tension d'écriture des cellules, ce qui augmentera également la contrainte sur la non-linéarité du sélecteur.

Un type de sélecteur considéré pour cette application est l'OTS pour *Ovonic Threshold Switch* (interrupteur à seuil ovonique) [93]. Il fonctionne de manière similaire à une mémoire à changement de phase. Le matériau entre les deux électrodes présente deux phases, l'une plus conductrice que l'autre. A la différence de la mémoire à changement de phase, la phase conductrice de l'OTS n'est pas stable en l'absence d'une excitation électrique. La phase conductrice se forme au-dessus d'une certaine tension et se maintient tant que la densité de courant ne diminue pas en dessous d'un seuil. Cette asymétrie entre les deux directions de transition entre phases permet à l'OTS de présenter une hystérésis. Ainsi, la tension de seuil n'a pas à être maintenue sur l'OTS pour la garder conductrice. Donc la chute de tension due au sélecteur peut être plus faible que ce seuil, à la différence de ce que cela serait avec une diode par exemple. La caractéristique de l'OTS est donc idéale pour servir de sélecteur back-end. Son principal problème est son manque de maturité.

VI.3.3 Cellules 1T1R

L'utilisation d'un sélecteur back-end serait idéale pour économiser de la surface active en permettant de mettre les circuits périphériques sous la matrice par exemple, ou pour envisager un assemblage mémoire 3D qui permettrait d'atteindre à terme des densités bien supérieures à n'importe quel assemblage 2D. Seulement, aujourd'hui, aucun sélecteur back-end ne propose une maturité permettant de réaliser des matrices de grande taille.

L'alternative choisie a été d'utiliser le transistor MOSFET pour ainsi former les cellules 1T1R discutées dans les précédents chapitres. Contrairement au sélecteur back-end qui repose sur l'utilisation d'une forte non-linéarité pour différencier la cellule sélectionnée des autres, la conductivité du transistor MOSFET est contrôlée par le potentiel de sa grille. Contrairement à une cellule 1S1R, la cellule 1T1R a trois pôles.

Ce pôle supplémentaire offre de nouvelles opportunités d'organisation de la matrice mémoire. On pourrait, par exemple, utiliser un adressage à trois signaux. Cela permettrait de pouvoir sélectionner une cellule parmi $N = n^3$ avec seulement $3\sqrt[3]{N} = 3n$ (au lieu de $2\sqrt{N} = 2n\sqrt{n}$) connexions[40]. Cependant, au-delà du problème géométrique de disposer un tel réseau tri-dimensionnel en deux dimensions, l'architecture à trois signaux présenterait le même problème de sélectivité approximative et requerrait l'utilisation d'un autre sélecteur pour limiter les courants de fuite et augmenter la sélectivité.

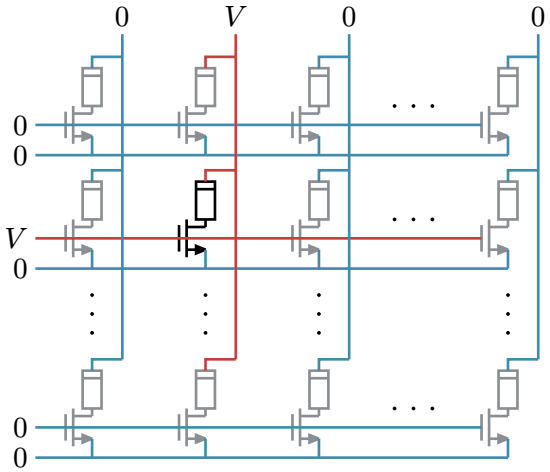


Figure 6.6 – Exemple de matrice 1T1R. Seule la cellule sélectionnée (en noir) ne peut conduire du courant.

[40] La dérivation d'une telle architecture reprend la même idée que celle à deux signaux décrite dans la partie VI.3.1, mais en utilisant trois partitions de la population au lieu de deux.

Nous restons donc sur une architecture à deux signaux. Plutôt que de sélectionner les cellules par leur deux pôles conducteurs, nous utilisons la grille du transistor ainsi que l'un des deux pôles conducteurs. L'autre pôle conducteur peut être mis en commun entre toutes les cellules sans diminuer la sélectivité. La Figure 6.6 illustre une telle disposition. En supposant le sélecteur parfait, il n'y a plus de problème de sélection approximative. En effet, la tension à appliquer à la cellule sélectionnée peut être mise sur le signal connecté sur son pôle conducteur adressable. Les autres signaux de ce groupe peuvent être gardés à la tension du pôle commun. Ainsi, une seule colonne de cellules subit la tension. La grille de la cellule sélectionnée est mise au potentiel permettant au transistor de conduire du courant. Cela rend les transistors de toute cette ligne conducteurs, cependant, seule la cellule sélectionnée a une différence de potentiel entre ces deux pôles conducteurs. De plus, les autres cellules voyant cette différence de potentiel n'ont pas leur transistor passant. Ainsi, il n'y a que la cellule sélectionnée qui est susceptible de conduire du courant. Nous sommes donc affranchis du problème rencontré précédemment.

En général, un transistor NMOS est utilisé comme sélecteur et connecté à l'élément mémoire OxRAM de telle sorte qu'il soit du côté du potentiel le plus faible en SET. Dans cette configuration, le transistor peut être utilisé en régime saturé en tant que limiteur de courant en réglant la tension entre sa grille et sa source qui est accessible. Il peut aussi n'être utilisé qu'en sélecteur, en régime linéaire, en fixant le potentiel de sa grille assez haut pour qu'il ne limite pas le courant. Pour une longueur de canal fixe, souvent choisie minimale, le courant maximal (fixé volontairement ou pas) pouvant être conduit par le transistor dépend à la fois de la tension grille/source et de sa largeur. Comme la tension est limitée, la largeur devient assez vite le principal paramètre définissant le courant pouvant être conduit par le transistor.

Dans cette architecture, le RESET présente un défi supplémentaire. En effet, en RESET, le pôle conducteur libre du transistor est au plus haut potentiel. Il agit donc comme son drain. Sa source est directement connectée à l'élément mémoire OxRAM. Ainsi, fixer le potentiel de grille ne permet pas de fixer la différence de potentiel grille/source qui détermine le courant maximal, mais surtout, la conductivité du transistor est fortement réduite. En effet, la majorité de la tension doit chuter sur l'élément mémoire OxRAM, cela signifie donc que le potentiel de la source du transistor est proche de celle du drain (celle que l'on fixe), ce qui n'est pas un problème en soi. Par contre, pour que le transistor reste conducteur, il est nécessaire d'avoir une différence de potentiel entre la grille et la source supérieure à la tension seuil du transistor. La conséquence est donc, qu'en RESET, la tension appliquée sur l'élément mémoire OxRAM ne peut dépasser la différence entre la tension maximale admissible de la technologie et la tension de seuil du transistor. Cela limite donc la finesse du transistor utilisé, et donc la longueur de son canal. Remplacer le transistor NMOS par un PMOS ou inverser la position du transistor par rapport à l'élément mémoire OxRAM ne résoudrait pas le problème, il ne ferait que l'inverser, résolvant le problème du RESET en le repoussant sur le SET. Une solution serait d'utiliser à la fois un NMOS et un PMOS pour former ce qui est généralement appelé une porte de transmission capable de conduire du courant aussi bien proche de la masse que proche de l'alimentation. Cependant cela demanderait une surface environ deux fois supérieure (voir plus si les transistors sont proches de leur taille minimum), ce qui est rédhibitoire d'un point de vue de la densité.

Pour que l'élément mémoire OxRAM soit fiable, il est nécessaire de l'écrire avec un minimum de courant durant le SET, et également au RESET puisque $I_c^{\text{eff}} \simeq I_{\text{RESET}}$ (voir Chapitre II.2.2). Le transistor doit être dimensionné pour permettre la circulation de ce courant. Ce courant minimal est assez indépendant de la taille de l'élément mémoire OxRAM du fait de sa conduction filamentaire. Du côté du transistor, une fois sa longueur de canal réduite au maximum, en respectant les contraintes de tension, ce courant minimum fixe un seuil de largeur de transistor en-dessous duquel nous ne pouvons pas descendre. La réduction de taille du transistor est donc limitée par ce besoin de tension et de

courant. Ainsi, la taille du transistor représente, dans cette architecture, une limite à l'augmentation de la densité. Elle est en pratique la première limite rencontrée.

VI.4 Conclusion

Ainsi, l'augmentation de la densité des mémoires OxRAM rencontre aujourd'hui différentes limites.

La réduction de la surface de l'élément mémoire OxRAM influence principalement la tension nécessaire à la formation du filament. Les courants mis en jeu lors de l'écriture ou l'effacement sont indépendants de cette surface. Cela pose un problème lors de l'augmentation de la densité, car tous les autres éléments du circuit voient leur capacité à conduire du courant diminuer avec leur taille.

Deux architectures sont mises en avant et présentent chacune leur problème. L'architecture crossbar souffre d'un problème de courant de fuite important. Des dispositifs back-end pourraient permettre de réduire ces courants de fuite, mais manquent généralement de maturité. Cette architecture ne permet donc pas de réaliser de grande matrice. L'architecture 1T1R profite de la maturité du transistor MOSFET et offre une meilleure sélectivité des éléments mémoires OxRAM, mais la densité quelle permet est limitée par la taille du transistor. L'asymétrie de la caractéristique du MOSFET réduit également la tension applicable sur l'élément mémoire OxRAM pendant l'effacement.

Ces deux voies d'intégration sont limitées en densité. La solution crossbar semble avoir le plus de perspectives d'évolution, mais pourra-t-on un jour fabriquer de façon fiable des sélecteurs présentant les caractéristiques requises pour l'intégration de cellules 1S1R à grande échelle ? Cela semble encore incertain. Quant aux matrices 1T1R, leur seule perspective est l'utilisation de technologie CMOS plus fine, offrant des densités de courant plus fortes. Cependant, les tensions mises en jeu lors de la formation du filament, mais également lors de l'écriture et de l'effacement doivent être réduites pour pouvoir être compatibles avec les nœuds CMOS les plus avancés.

Par conséquent, dans le prochain chapitre, nous étudierons une solution alternative, à mi-chemin entre les crossbar et les matrices 1T1R pour tenter de repousser ces limites de densité.

Bibliographie

[60] J. Sandrini *et al.*, « OxRAM for embedded solutions on advanced node: scaling perspectives considering statistical reliability and design constraints », in *2 019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 30. doi: 10.1109/IEDM19573.2019.8993484.

[76] M. Barlas *et al.*, « Improvement of HfO2 based RRAM array performances by local Si implantation », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 14. doi: 10.1109/IEDM.2017.8268392.

[77] L. Grenouillet *et al.*, « 16kbit 1T1R OxRAM arrays embedded in 28 nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors », in *2 021 IEEE International Memory Workshop (IMW)*, mai 2021, p. 1-4. doi: 10.1109/IMW51353.2021.9439607.

[88] « Cypress eCT Flash: The Most Scalable High-Performance, High-Reliability Embedded Flash Solution for Next-Generation MCUs », in *Infineon Product Brief "eCT Flash"*, 2021.

[89] J. Cooke, « NAND Flash 101: An Introduction to NAND Flash and How to Design It in to Your Next Product », in *Micron Technical Note 29 19 (TN-29-19)*, 2016.

[90] A. Lu, X. Peng, W. Li, H. Jiang, et S. Yu, « NeuroSim Simulator for Compute-in-Memory Hardware Accelerator: Validation and Benchmark », *Frontiers in Artificial Intelligence*, vol. 4, p. , 2021, doi: 10.3389/frai.2021.659060.

[91] T. Liu, S. Zhu, M. H. White, A. Salemi, D. Sheridan, et A. K. Agarwal, « Time-Dependent Dielectric Breakdown of Commercial 1.2 kV 4H-SiC Power MOSFETs », *IEEE Journal of the Electron Devices Society*, vol. 9, p. 633-639, 2021, doi: 10.1109/JEDS.2021.3091898.

[92] H.-X. Zheng *et al.*, « Reducing Forming Voltage by Applying Bipolar Incremental Step Pulse Programming in a 1T1R Structure Resistance Random Access Memory », *IEEE Electron Device Letters*, vol. 39, nº 6, p. 815-818, juin 2018, doi: 10.1109/LED.2018.2831708.

[93] D. A. Robayo *et al.*, « Integration of OTS based back-end selector with HfO2 OxRAM for crossbar arrays », in *2 019 IEEE 11th International Memory Workshop (IMW)*, mai 2019, p. 1-4. doi: 10.1109/IMW.2019.8739746.

Chapitre VII : Une solution au problème de densité : les matrices 1TnR ?

VII.1 Introduction

Nous avons vu dans le chapitre précédent que deux architectures sont particulièrement présentes dans la littérature, l'architecture crossbar et l'architecture 1T1R. Néanmoins, nous avons aussi identifié leur limite concernant la densité qu'elles permettent.

Dans ce chapitre, nous étudions une architecture que l'on pourrait qualifier d'intermédiaire entre crossbar et 1T1R. On la nomme souvent 1TnR car elle n'utiliser qu'un transistor pour n éléments mémoires OxRAM. Cela permet donc en théorie d'augmenter la densité. Cependant ce partage conduit inévitablement à l'apparition de courant parasite comme pour la crossbar.

Cette architecture est présente dans la littérature [94] [95] [96] [97] [98] [99], mais il n'est pas facile d'évaluer les compromis mis en jeu au travers des informations et données fournies par celle-ci. Nous nous engageons donc dans la voie de produire notre propre circuit expérimentant cette architecture, le circuit *OX1TnR*. L'objectif de ce circuit est de fournir un prototype implémentant la nouvelle architecture 1TnR[41] décrit dans ce chapitre.

Ce chapitre décrit l'architecture 1TnR développée ainsi que les raisons techniques qui nous ont conduits à celle-ci. Nous décrivons d'un point de vue théorique les méthodes d'écriture et d'effacement utilisables et essayons de donner une méthode pour leur optimisation. Nous discutons ensuite des techniques de lecture envisageables. Pour finir, nous discutons des détails d'implémentation du circuit, et plus particulièrement du dessins des masques du cœur de la matrice mémoire.

VII.2 Matrice 1TnR

VII.2.1 Cellule 1TnR

Une cellule 1TnR est composée de n éléments mémoires OxRAM et d'un transistor. Ils sont connectés ensemble de telle sorte à ce que chaque élément mémoire OxRAM ait un pôle libre et un pôle connecté au drain du transistor, comme illustré sur la Figure 7.1. Dans la suite, les lignes connectées au pôle libre seront dénotées « BL », celle connectée à la source du transistor « SL » et celle connectée à la grille du transistor « GL ». Les schémas utiliseront des cellules 1T2R comme exemple, mais des cellules avec plus d'éléments mémoires OxRAM pourraient être utilisées de manière équivalente.

[41] L'architecture décrite a fait l'objet d'une demande de dépôt de brevet T. Bauvent, G. Pillonnet, G. Molas, « Dispositif mémoire à matrice de cellules mémoires », FR2412063 – A1.

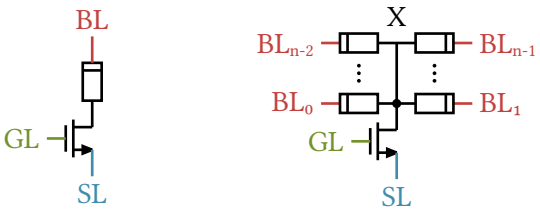
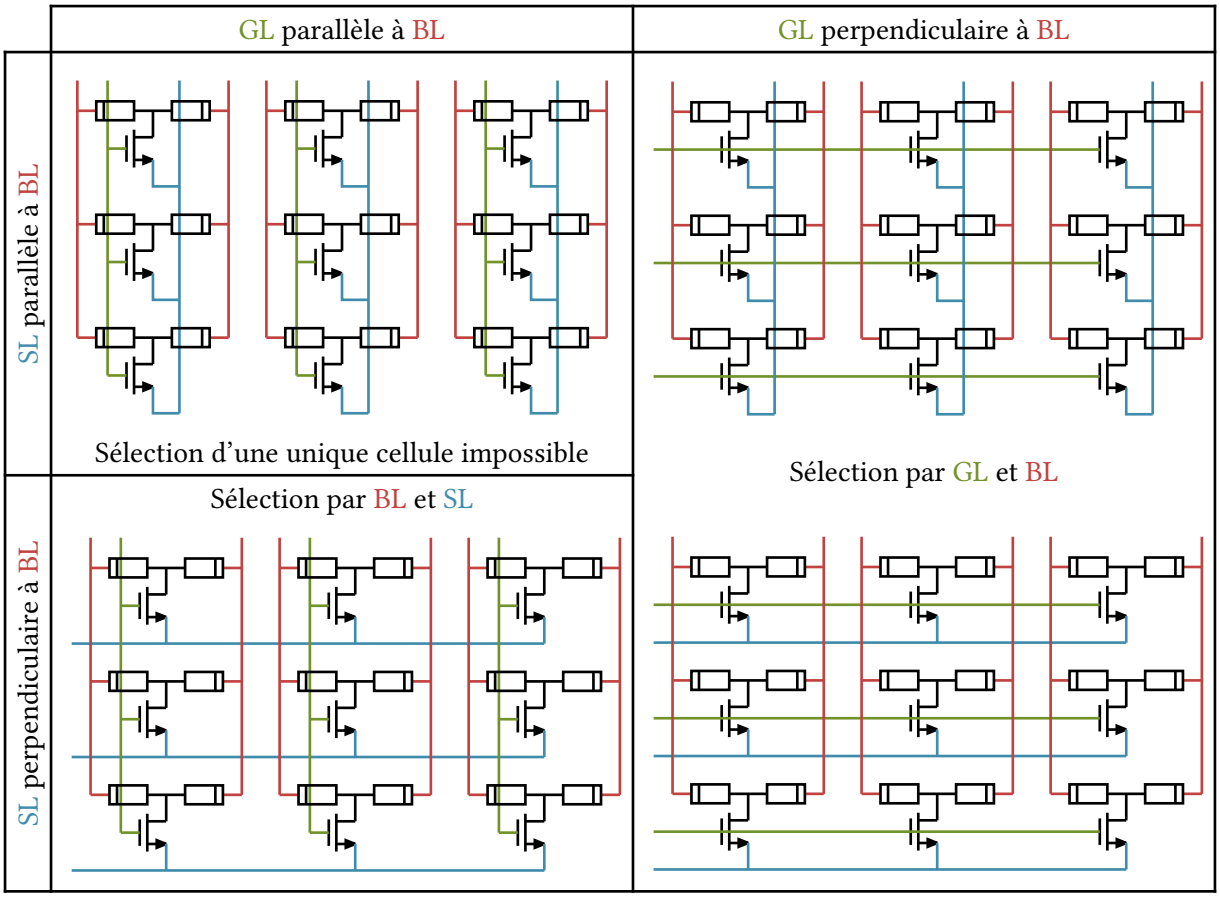


Figure 7.1 – Comparaison entre une cellule 1T1R vue dans les précédents chapitres et une cellule 1TnR. Pour la cellule 1TnR, n éléments mémoires OxRAM sont connectés au même transistor, et n BL sont disponibles.

Tableau 7.1 – Possibilités d'orientation des connexions au sein de la matrice



VII.2.2 Intégration sous forme de matrice

Sans *a priori*, on a quatre possibilités pour connecter nos trois types de terminaux de façon régulière en suivant le schéma de la matrice 2D présenté au chapitre précédent. Le Tableau 7.1 résume ces quatre possibilités en fixant, sans perte de généralité, les BL verticales. Pour permettre la sélection de cellules, il est nécessaire d'avoir au moins un type de ligne perpendiculaire aux BL. Donc la première organisation avec GL et SL parallèles aux BL, est inutilisable. Lorsque GL est perpendiculaire aux BL, la sélection se fait aux travers de GL et BL. La direction des SL est sans grande importance, elles pourraient même, en principe, être toutes connectées entre elles. Ces deux organisations sont directement dérivées de celles des matrices 1T1R en remplaçant la cellule 1T1R par une cellule 1TnR. On l'appellera dans la suite GL⊥ La dernière option requiert un adressage par SL et BL comme dans l'architecture crossbar. On l'appellera SL⊥ dans la suite.

On commence d'abord par identifier les différences entre les dispositions GL⊥ et SL⊥. Pour ce faire, on s'intéresse à l'écriture. La Figure 7.2 montre une manière d'effectuer une écriture sur plusieurs

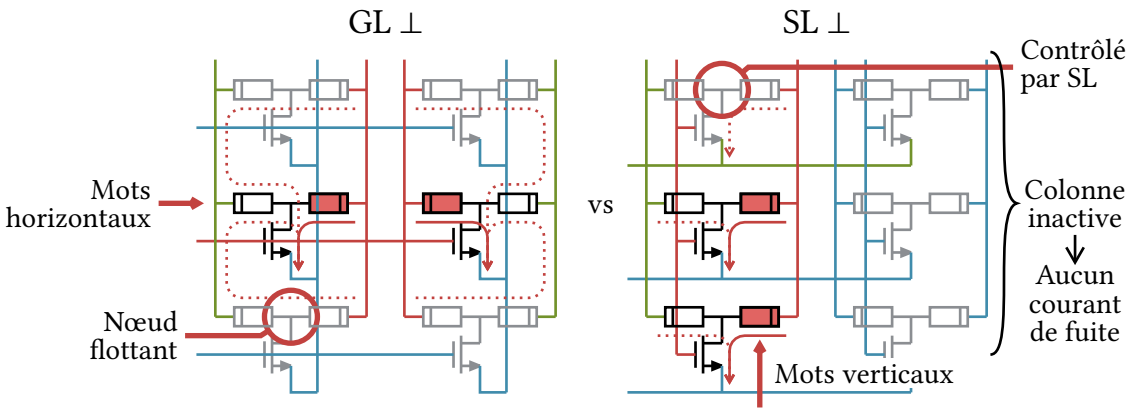


Figure 7.2 – Illustration des différences entre GL⊥ (gauche) et SL⊥ (droite). Les lignes sont colorées selon leur potentiel, rouge: V, vert: $\frac{V}{2}$ et bleu: 0 V. Ces choix de potentiel correspondent à une stratégie d'écriture des éléments mémoires OxRAM représentés avec un fond rouge. Les flèches rouges en trait plein représentent les courants utiles à l'écriture et les traits en pointillés les courants de fuite.

Tableau 7.2 – Comparaison entre les architectures GL⊥ et SL⊥

Quantité	GL⊥	SL⊥
Nœuds internes	Flottants sauf celui sélectionné	Tous contrôlables par les SL
Accessible en parallèle	$N \times$ nombre de colonnes actives[α]	M[β]
Courant parasite	$\propto$ nombre de colonnes actives	indépendant du nombre d'accès

[α] N = nombre de dispositifs par sélecteur
[β] M = nombre de sélecteurs par colonne = nombre de lignes SL

éléments mémoires OxRAM en même temps dans les deux architectures. Nous commençons d'abord par remarquer les différences qui sont résumées dans le Tableau 7.2.

1. **Direction des mots binaires**
   La direction des mots est la direction dans laquelle il est possible d'effectuer des écritures simultanées de plusieurs éléments mémoires OxRAM à la fois. Comme l'écriture d'un élément mémoire OxRAM requiert un transistor passant, la direction des mots est identique à celle de GL. Donc dans le cas de GL⊥, elle est horizontale, et les bits du mot à écrire sont sélectionnés par les BL. Dans le cas de SL⊥, elle est verticale, et les bits du mot à écrire sont sélectionnés par les SL. Cette différence a des conséquences intéressantes dès lors que l'on considère l'écriture simultanée de plusieurs éléments.
2. **Quantité de chemin de fuite**
   Dans l'architecture GL⊥, la sélection des bits est effectuée par les BL. En fonction de la distribution de ces bits, plusieurs colonnes de cellules seront potentiellement mises en jeu. La différence de tension entre les BL d'une même colonne de cellules sont à l'origine des courants de fuite dans les cellules non sélectionnées. Dans le pire des cas, où au moins un bit de chaque colonne doit être écrit, l'intégralité de la matrice participera à la conduction du courant utile, mais également du courant de fuite.
   Au contraire, dans l'architecture SL⊥, qui sélectionne les bits à écrire par les SL, seule une colonne de cellules est active à la fois. Ainsi, au plus une colonne participera à la conduction de courant de fuite. Les courants de fuite sont donc indépendants du nombre de colonnes. Ainsi au lieu d'augmenter, au pire des cas, avec le nombre d'éléments mémoires OxRAM, comme dans l'organisation GL⊥, ils ne augmentent qu'avec la racine carrée du nombre d'éléments mémoires OxRAM.
3. **Contrôle du potentiel des nœuds**
   Le potentiel du nœud interne à la cellule 1TnR, celui entre le drain du transistor et les éléments

mémoires OxRAM dénotés X sur la Figure 7.1, est important car il influence la différence de potentiel perçue par chaque élément mémoire OxRAM.
Dans l'architecture GL⊥, pour les cellules n'étant pas sur la GL sélectionnée, le potentiel de leur nœud interne est mal défini. En effet, comme le transistor ne conduit pas, il est entièrement défini par le rapport de conductance des n éléments mémoires OxRAM de la cellule ainsi que du potentiel de leur terminal libre. Comme la conductance des n éléments mémoires OxRAM est supposée inconnue[42], on ne peut donc seulement en déduire que le potentiel du nœud est entre les potentiels des BL.
Au contraire, l'architecture SL⊥ permet de contrôler le potentiel de ces mêmes nœuds internes au travers des SL et des transistors (qui sont passants car sur la GL active). Les colonnes non sélectionnées, celles n'ayant pas leur transistor passant, ne présentent pas le même problème car toutes les BL de la colonne sont au même potentiel, ce qui définit parfaitement le potentiel des nœuds internes indépendamment de l'état des éléments mémoires OxRAM.

Notons maintenant les similarités entre GL⊥ et SL⊥ :

1. Plus le nombre n d'éléments mémoires OxRAM par transistor est élevé, moins la SL au travers du transistor contrôle le potentiel du nœud interne à la cellule. En effet, le potentiel du nœud interne est égal à la moyenne des potentiels des BL et de la SL pondérée par la conductance des éléments mémoires OxRAM pour les BL et du transistor pour la SL. Idéalement, on souhaiterait que le potentiel du nœud interne soit égal à celui de la SL. Cependant, la conductance du transistor est finie. Plus elle est proche de la somme de conductance des n éléments mémoires OxRAM, plus les potentiels des BL influenceront le potentiel du nœud interne. Le nombre n d'éléments mémoires OxRAM par transistor est donc limité par le rapport de conductance du transistor et le régime de conductance des éléments mémoires OxRAM.
2. Nous avons vu que l'écriture simultanée de plusieurs éléments mémoires OxRAM représentait une différence importante entre GL⊥ et SL⊥. Néanmoins, le cas de l'écriture d'un unique élément mémoire OxRAM est quant à lui similaire. Dans les deux cas, les courants parasites sont proportionnels au nombre n d'éléments mémoires OxRAM par transistor ainsi que du nombre de SL par colonne. De plus, dans ce cas, SL⊥ peut être rendu équivalent à GL⊥ en laissant les SL non sélectionnées flottantes. De ce point de vue, SL⊥ est une solution strictement supérieure à GL⊥ car elle présente un degré de liberté supplémentaire utilisable pour optimiser le fonctionnement.

Pour ces raisons, nous nous intéressons dans la suite à l'architecture SL⊥ que nous avons identifiée comme supérieure à GL⊥.

VII.2.3 Technique d'écriture

Nous avons déjà discuté d'une méthode d'écriture des éléments mémoires OxRAM pour comparer les architectures GL⊥ et SL⊥. Cela nous a permis de choisir SL⊥. Nous rentrons maintenant dans plus de détails. Nous commençons par essayer de dériver les différentes méthodes envisageables en établissant les contraintes à respecter.

Pour écrire un élément mémoire OxRAM, il est nécessaire de lui appliquer une tension suffisamment grande. Toutefois, il ne faut par ailleurs pas dépasser une certaine tension sur l'ensemble des autres dispositifs que l'on risquerait, sinon, d'écrire involontairement. Notons que cette tension seuil entre écriture et non-écriture dépend du temps pendant lequel l'élément est exposé à celle-ci. De plus elle ne

[42]Supposer connaître la conductance des éléments mémoires OxRAM revient à supposer connaître a minima le contenu de la mémoire. On admettra qu'une mémoire dont le contenu doit être connu pour permettre son utilisation n'est pas vraiment pratique. Pour les opérations d'écriture, on pourrait considérer effectuer une lecture au préalable pour obtenir cette information, mais cela serait contraignant, on ne souhaite donc pas être dépendant de la connaissance de l'état de la mémoire. Également, on pourrait considérer limiter les états possibles, en interdisant que tous les éléments mémoires OxRAM soient en HRS par exemple. Cela permettrait d'avoir une information partielle sur l'état de la conductance des éléments mémoires OxRAM.

reflète en réalité qu'une certaine probabilité d'écriture de l'élément. Il donc préférable de considérer deux seuils, l'un assurant l'écriture presque certainement et l'autre assurant la non-écriture presque certainement.

De façon générale, considérons que l'on applique une tension V sur l'élément mémoire OxRAM à écrire en appliquant un potentiel V sur la BL correspondante et 0 V sur la SL. Il nous reste alors à fixer les potentiels des autre BL et SL. Comme nous l'avons vu, l'architecture SL⊥ permet de considérer les différentes colonnes de cellules de façon indépendante, comme dans le cas d'une matrice 1T1R. Nous ne considérons donc que la colonne sélectionnée dans la suite. Les potentiels des BL appartenant aux colonnes non sélectionnées peuvent être fixés à un potentiel commun pour prévenir de toute perturbation sur les éléments mémoires OxRAM de ces colonnes. Pour ce qui est de la colonne sélectionnée, il est nécessaire de trouver des potentiels adaptés. En l'absence d'un *a priori* sur l'état de conductance des éléments mémoires OxRAM[43], il n'y a aucune raison d'appliquer un potentiel différent sur différentes BL non sélectionnées ou sur différentes SL non sélectionnées.

Nous n'avons donc que trois degrés de liberté à fixer, le potentiel des BL non sélectionnées V_{BL}^u, le potentiel des SL non sélectionnées V_{SL}^u, et la différence de potentiel V appliquée à l'élément mémoire OxRAM à écrire. En notant V_L la tension limite à ne pas dépasser pour s'assurer de ne pas écrire un élément mémoire OxRAM, V_{X}^s le potentiel du nœud interne X de la cellule contenant l'élément mémoire OxRAM à écrire et V_{X}^u celui des nœuds internes aux autres cellules, on obtient que :

1. $\mid V_{\mathrm{BL}}^u - V_{\mathrm{X}}^s \mid \leq V_L$ pour s'assurer de ne pas écrire les éléments mémoires OxRAM partageant le transistor de l'élément mémoire OxRAM à écrire;
2. $\mid V_{\mathrm{BL}}^u - V_{\mathrm{X}}^u \mid \leq V_L$ pour s'assurer de ne pas écrire les éléments mémoires OxRAM ne partageant ni le transistor, ni la BL de l'élément mémoire OxRAM à écrire.
3. $\mid V - V_{\mathrm{X}}^u \mid \leq V_L$ pour s'assurer de ne pas écrire les éléments mémoires OxRAM partageant la BL de l'élément mémoire OxRAM à écrire;

En considérant en premier lieu la conductance des transistors infinie, on a $V_{\mathrm{X}}^s = 0\,\mathrm{V}$ et $V_{\mathrm{X}}^u = V_{\mathrm{SL}}^u$. Les inégalités précédentes sont alors réduites à $\mid V_{\mathrm{BL}}^u \mid \leq V_L$, $\mid V_{\mathrm{SL}}^u - V_{\mathrm{BL}}^u \mid \leq V_L$ et $\mid V - V_{\mathrm{SL}}^u \mid \leq V_L$. On a donc dans tous les cas $|V| < 3V_L$[44]. Sans respecter cette condition, il ne sera pas possible d'écrire un élément mémoire OxRAM sans risquer d'en écrire un autre.

On remarque ensuite que chaque différence de potentiel est associée à un type de courant :

- V est la tension d'écriture de l'élément mémoire OxRAM et est associée au courant utile d'écriture;
- $V - V_{\mathrm{SL}}^u$ est liée au courant parasite sur la BL de l'élément mémoire OxRAM à écrire;
- V_{BL}^u est liée au courant parasite passant dans le transistor et la SL de la cellule contenant l'élément mémoire OxRAM à écrire;
- $V_{\mathrm{SL}}^u - V_{\mathrm{BL}}^u$ est liée au courant parasite passant dans les éléments ne partageant ni le transistor, ni la BL de l'élément mémoire OxRAM à écrire.

Ainsi, on peut adopter différentes stratégies chacune visant à optimiser un paramètre différent. Quatre stratégies sont illustrées sur la Figure 7.3.

1. **Maximiser la tension V**

   Pour permettre d'avoir la plus forte tension possible sur l'élément mémoire OxRAM à écrire, c'est-à-dire V maximale, il est nécessaire d'espacer le plus possible les potentiels entre eux, ce qui donne pour un SET $V_{\mathrm{BL}}^u = V_L$, $V_{\mathrm{SL}}^u = 2V_L$ et $V = 3V_L$.

[43] Cette hypothèse de ne pas avoir d'*a priori* permet une plus grande généralité. Dans l'éventualité où la démarche décrite dans ce manuscrit mènerait à la conclusion qu'une méthode d'écriture convenable ne peut exister, cette hypothèse pourra être remise en question pour relaxer le problème et espérer trouver une solution.

[44] En remarquant que $V = V - V_{\mathrm{SL}}^u + V_{\mathrm{SL}}^u - V_{\mathrm{BL}}^u + V_{\mathrm{BL}}^u$, puis en sommant les trois inégalités $-3V_L \leq V \leq 3V_L$.

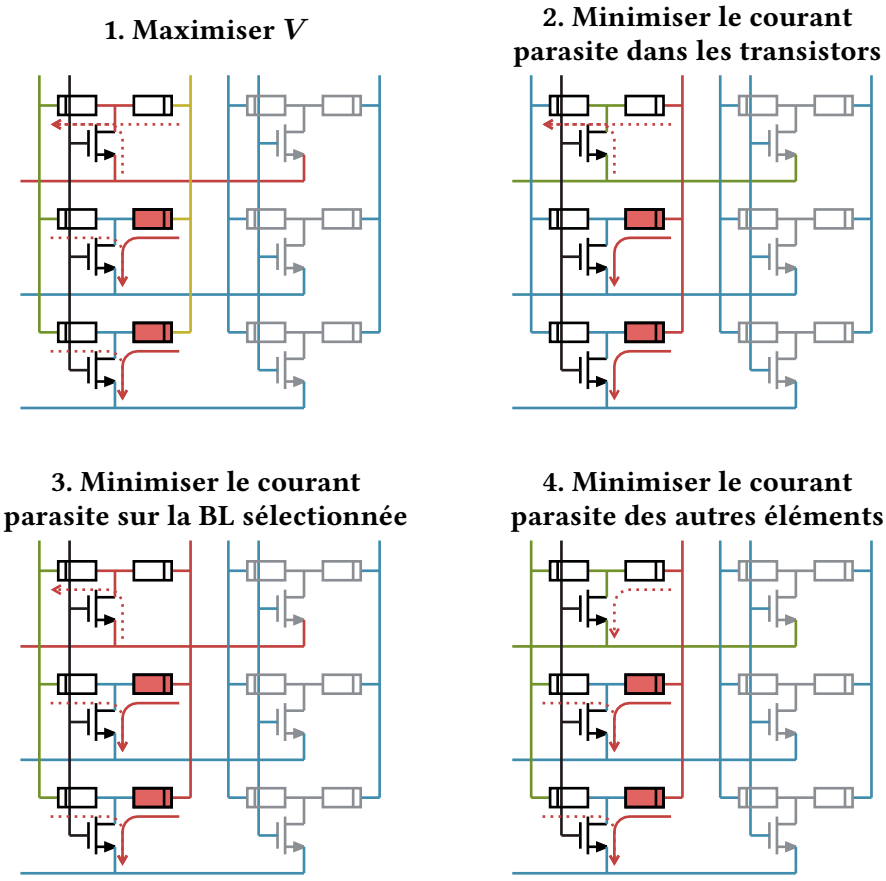


Figure 7.3 – Quatre stratégies de SET optimisant chacune un paramètre différent. Les éléments mémoires OxRAM que l'on essaie d'écrire ont un fond rouge. Les couleurs des lignes correspondent à leur potentiel, 0 V (bleu), V_L (vert), $2V_L$ (rouge), $3V_L$ (jaune). Les flèches rouges représentent les courants utiles (en trait plein) et parasites (en pointillé).

2. **Minimiser le courant parasite dans le transistor et la SL de l'élément sélectionné**
   On peut vouloir réduire le courant parasite passant dans le transistor et la SL de l'élément mémoire OxRAM à écrire. Cela peut permettre de réduire le courant que le transistor doit être capable de conduire, et donc sa surface, ou cela peut également permettre de mesurer le courant passant dans l'élément à écrire par la SL. Ce courant est dû à la différence de potentiel entre V^u_{BL} et V^s_{SL}. Pour le réduire, on peut fixer cette différence à zéro. On a alors pour un SET $V^u_{\mathrm{BL}} = 0$. Si l'on essaie ensuite de maximiser V avec cette contrainte, on obtient cette fois, $V^u_{\mathrm{BL}} = 0$, $V^u_{\mathrm{SL}} = V_L$ et $V = 2V_L$.
3. **Minimiser le courant parasite dans la BL de l'élément sélectionné**
   On peut vouloir réduire le courant parasite passant dans la BL sélectionnée, par exemple pour réduire le courant que la périphérie doit être capable de fournir, en particulier lors de l'écriture de plusieurs éléments en simultané. Ce courant est dû à la différence de potentiel entre V^s_{BL} et V^u_{SL}. En fixant cette différence à zéro puis en maximisant V, on obtient alors $V^u_{\mathrm{BL}} = V_L$, $V^u_{\mathrm{SL}} = 2V_L$ $V = 2V_L$.
4. **Minimiser le courant parasite dans les autres BL**
   Enfin, on peut vouloir réduire le courant parasite passant dans les éléments mémoires OxRAM qui ne sont connectés ni à la BL ni à la SL sélectionnées, qui sont souvent les plus nombreux et représentent donc la majorité du courant de fuite. Ce courant est dû à la différence de potentiel entre V^u_{SL} et V^u_{BL}. En fixant cette différence à zéro puis en maximisant V, on obtient $V^u_{\mathrm{BL}} = V_L$, $V^u_{\mathrm{SL}} = V_L$ et $V = 2V_L$.

Ces stratégies peuvent être mélangées pour répondre aux contraintes d'une implémentation particulière, cependant, il existe un compromis fort entre la tension V atteignable et la réduction des différents courants parasites. Le RESET peut être réalisé de manière similaire, en prenant l'opposé de tous les potentiels.

Pour l'instant, la conductance finie du transistor n'a pas été prise en compte. En réalité, le potentiel des nœuds internes aux cellules n'est pas celui des SL, mais une moyenne entre celui des SL et des BL pondérée par la conductance des transistors et des éléments mémoires OxRAM. Sans *a priori* sur l'état des éléments mémoires OxRAM, le potentiel des nœuds internes est donc indéterminé. En connaissant la plage de conductance des éléments mémoires OxRAM ainsi que la caractéristique du transistor, il est possible de calculer l'intervalle dans lequel sont les potentiels internes. Il est alors nécessaire de prendre une certaine marge sur les tensions que l'on a précédemment déterminées pour assurer que la tension sur chaque élément mémoire OxRAM soit inférieure à V_L (sauf sur celles que l'on souhaite écrire) quel que soit l'état de conductance des éléments mémoires OxRAM alentour.

Dans les cellules sélectionnées, le potentiel de la SL est toujours le plus faible (0 V). Du fait de la conductance finie du transistor, le potentiel du nœud interne V_{X}^s est supérieur à 0 V. La contrainte $V_{\mathrm{BL}}^u \leq V_L$ satisfait donc bien $\mid V_{\mathrm{BL}}^u - V_{\mathrm{X}}^s \mid \leq V_L$. Par contre la tension effective sur l'élément mémoire OxRAM à écrire s'en voit réduite. Dans les cellules non sélectionnées, le potentiel du nœud interne V_{X}^u, par rapport à celui de V_{SL}^u, peut aussi bien être tiré vers le haut par la BL sélectionnée que vers le bas par les BL non sélectionnées.

L'augmentation de V_{X}^u affecte la contrainte $\mid V_{\mathrm{X}}^u - V_{\mathrm{BL}}^u \mid \leq V_L$ concernant les éléments mémoires OxRAM n'étant ni sur la BL, ni sur les SL sélectionnées. Dans l'hypothèse où $V_{\mathrm{BL}}^u < V_{\mathrm{SL}}^u$, la valeur maximale atteignable par V_{X}^u est $V_{\mathrm{SL}}^u + \frac{\gamma_{\max}}{\gamma_0+\gamma_{\max}}(V - V_{\mathrm{SL}}^u)$ avec $\gamma_{\max}$ la conductance maximale d'un élément mémoire OxRAM (plus particulièrement, de celui sur la BL sélectionnée) et γ_0 la conductance du transistor. Cette limite est atteinte lorsque tous les éléments mémoires OxRAM de la cellule ont une conductance nulle, sauf celle sur la BL sélectionnée qui a comme conductance $\gamma_{\max}$.

La diminution de V_{X}^u affecte quant à elle la contrainte $\mid V - V_{\mathrm{X}}^u \mid \leq V_L$ concernant les éléments mémoires OxRAM sur la BL sélectionnée. Avec toujours $V_{\mathrm{BL}}^u < V_{\mathrm{SL}}^u$, la valeur minimale atteignable par V_{X}^u est $V_{\mathrm{SL}}^u - \frac{(n-1)\gamma_{\max}}{\gamma_0+(n-1)\gamma_{\max}}(V_{\mathrm{SL}}^u - V_{\mathrm{BL}}^u)$ avec $\gamma_{\max}$ la conductance maximale d'un élément mémoire OxRAM (plus particulièrement, de ceux n'étant pas sur la BL sélectionnée), γ_0 la conductance du transistor et n le nombre d'éléments mémoires OxRAM dans une cellule. Cette limite est atteinte lorsque tous les éléments mémoires OxRAM de la cellule ont comme conductance $\gamma_{\max}$, sauf celle sur la BL sélectionnée qui aurait une conductance nulle.

On peut en déduire que les potentiels V et V_{SL}^u sont tels que décrits par (7.1) et (7.2).

$$V_{\mathrm{SL}}^u - V_{\mathrm{BL}}^u \leq \frac{\gamma_0 + (n-1)\gamma_{\max}}{\gamma_0 + n\gamma_{\max}} V_L \tag{7.1}$$

$$V - V_{\mathrm{SL}}^u \leq \frac{\gamma_0 + \gamma_{\max}}{\gamma_0 + n\gamma_{\max}} V_L \tag{7.2}$$

Appliqué au cas où l'on cherche a avoir V maximal, on obtient (7.3). La tension maximale applicable est donc réduite par la conductance finie du transistor, et l'est d'autant plus que la conductance totale des éléments mémoires OxRAM connectés, $n\gamma_{\max}$, s'en rapproche.

$$V \leq \left(3 - \frac{n\gamma_{\max}}{\gamma_0 + n\gamma_{\max}}\right) V_L \tag{7.3}$$

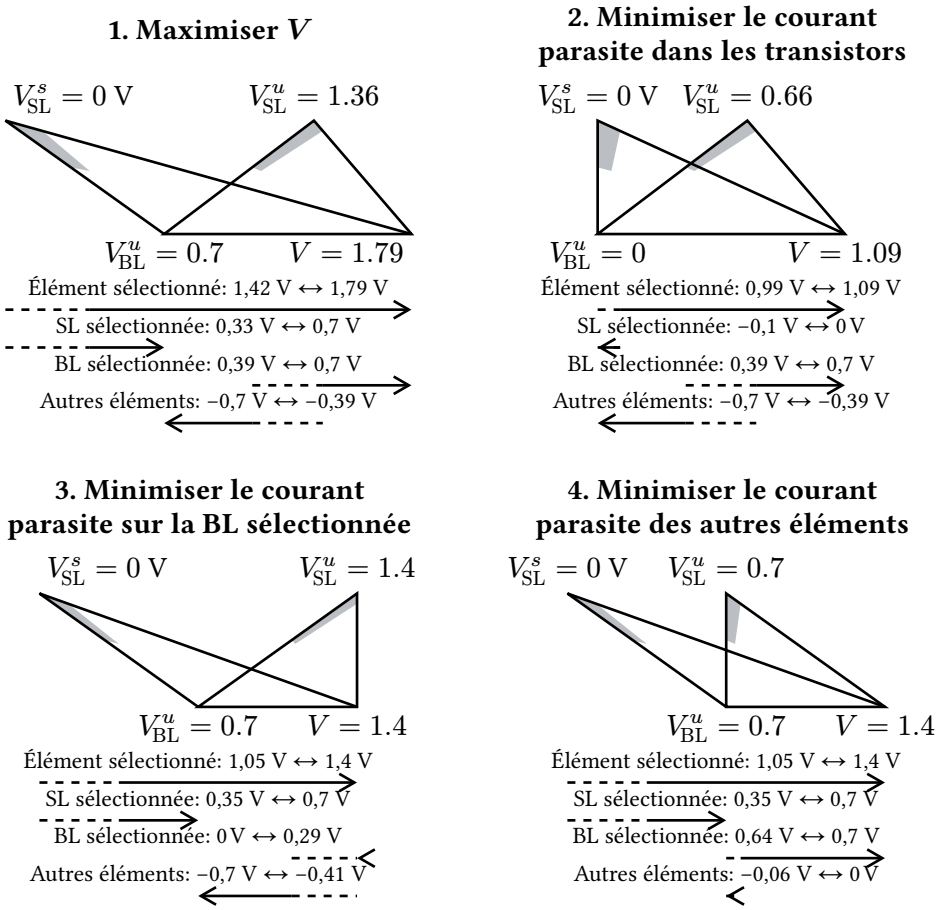


Figure 7.4 – Représentation permettant la détermination des potentiels V_{X}^s et V_{X}^u. La zone grise représente leur lieu. L'axe des abscisses correspond au potentiel. Les potentiels ont été choisis selon les quatre stratégies de la Figure 7.3, et en appliquant (7.1) et (7.2) lorsque nécessaire en considérant $V_L = 0{,}7\,\mathrm{V}$, que les éléments mémoires OxRAM ont une conductance comprise entre 0 µS et 200 µS, que le transistor a une conductance constante de 2000 µS et que $n = 8$.

Dans le cas réduisant[45] le courant passant dans le transistor de la cellule sélectionnée où $V_{\mathrm{BL}}^u = 0\,\mathrm{V}$, on obtient la même tension V maximale avec un V_L de moins, donc un peu moins de $2V_L$.

Dans le cas réduisant[45] le courant sur la BL sélectionnée où $V_{\mathrm{SL}}^u = V$, on a $V = V_{\mathrm{SL}}^u = V_{\mathrm{BL}}^u + V_L = 2V_L$.

Enfin, dans le cas réduisant[45] le courant dans les éléments mémoires OxRAM ne partageant ni la BL ni une des SL sélectionnées où $V_{\mathrm{SL}}^u = V_{\mathrm{BL}}^u$, on a $V = V_{\mathrm{SL}}^u + V_L = 2V_L$.

De façon moins analytique et plus géométrique, il est possible de visualiser ces différents compromis à l'aide de deux triangles. L'idée est de représenter le lieux de V_{X}^s, respectivement V_{X}^u, en coordonnées barycentriques dans un triangle dont les trois sommets correspondent à V_{BL}^u, V et 0 V, respectivement V_{SL}^u. On peut ensuite placer les quatre points sur une abscisse représentant leur potentiel. L'intervalle de potentiel possible pour V_{X}^s et V_{X}^u peut alors être obtenu par projection de leur lieu sur l'axe des abscisses. Cette procédure a l'intérêt de faciliter la détermination des configurations de conductance des éléments mémoires OxRAM à l'origine des bornes des intervalles recherchés. La Figure 7.4 illustre

[45] Réduit et non pas minimise ou annule car $V_{\mathrm{SL}} \neq V_{\mathrm{X}}$. Il y a donc tout de même un courant dans les éléments mémoires OxRAM visés. A cause de l'incertitude sur V_{X}, on ne peut pas annuler le courant sans connaître l'état des éléments.

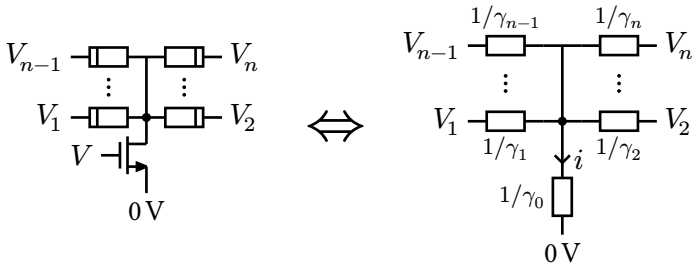


Figure 7.5 – Cellule 1TnR sélectionnée (le sélecteur est passant) vue comme un réseau de résistance. Le courant i peut être observé à travers la SL.

un exemple d'une telle représentation utilisant les potentiels définis à l'aide de (7.1) et (7.2) pour les quatre stratégies de la Figure 7.3.

VII.2.4 Technique de lecture

a) Lecture absolue du niveau de conductance

Dans l'architecture 1T1R, la lecture est réalisée en imposant une faible tension (par exemple 0,2 V) sur l'élément mémoire OxRAM à lire et en mesurant le courant la traversant. On peut essayer de faire la même chose dans notre architecture 1TnR. Pour ce faire, considérons une cellule 1TnR sélectionnée comme un réseau de résistance, comme représentée sur la Figure 7.5. On peut mesurer i passant dans le sélecteur en mesurant le courant passant dans la SL, comme on pourrait le faire sur une cellule 1T1R. Cependant, ici, on ne peut pas directement déduire la conductance d'un élément mémoire OxRAM en particulier à partir d'une unique mesure de i. En effet, on a $i = \gamma_0 \frac{\sum_{i=1}^{n} \gamma_i V_i}{\sum_{i=0}^{n} \gamma_i}$. En normalisant par $\gamma_0 V$ avec V la tension de lecture typique, on obtient $\frac{i}{\gamma_0 V} = \frac{\sum_{i=1}^{n} x_i S_i}{\sum_{i=0}^{n} x_i}$ avec $x_i = \frac{\gamma_i}{\gamma_0}$ et $S_i = \frac{V_i}{V}$. Supposons que l'on effectue n mesures correspondant chacune à une situation différente où l'une des tension V_i est égale à V pendant que les autres sont nulles. En notant M_j le courant normalisé observé pour chacune, on obtient $M_j = \frac{x_j}{\sum_{i=0}^{n} x_i}$. Remarquons que $\sum_{i=1}^{n} M_i + \frac{1}{\sum_{i=0}^{n} x_i} = 1$ et donc que $\sum_{i=0}^{n} x_i = \frac{1}{1-\sum_{i=1}^{n} M_i}$. Cela nous permet d'inverser la relation et d'obtenir $x_j = \frac{M_j}{1-\sum_{i=1}^{n} M_i}$.

En considérant que γ_0 est connu, on peut donc déduire la valeur des γ_i. Cependant, nous remarquons qu'il est nécessaire d'avoir effectué les n mesures, même si un seul γ_i est recherché. Dans la mesure où tous les éléments mémoires OxRAM de la cellule doivent être lus, au calcul supplémentaire près, l'opération demande autant d'étapes que sur n cellules 1T1R.

Dans l'hypothèse où la somme des conductances des éléments mémoires OxRAM est négligeable devant la conductance du sélecteur γ_0, la relation peut être simplifiée en $x_i = M_i$ en remarquant que $1 - \sum_{i=1}^{n} M_i \sim 1$. En effectuant cette approximation, la valeur de conductance de chaque élément mémoire OxRAM peut alors être obtenue indépendamment en effectuant une unique mesure.

b) Lecture relative

L'utilisation des éléments mémoires OxRAM en tant que mémoire ne requiert pas la mesure de la conductance de chaque élément mémoire OxRAM. Être capable de distinguer l'état de chaque élément mémoire OxRAM suffit à pouvoir lire l'information stockée. En imposant une tension V_i sur le terminal libre de chaque élément mémoire OxRAM au sein d'une cellule 1TnR et un courant nul dans le transistor (en présentant une forte impédance sur SL), l'état de conductance relatif des éléments mémoires OxRAM, noté a_i, peut être déterminé. En effet, le potentiel du terminal commun à tous les éléments mémoires OxRAM de la cellule V_0, qui est également celui de la SL, est égal à la moyenne des potentiels pondérée par la conductance de chaque élément mémoire OxRAM comme exprimé par (7.4). La valeur d'un a_i peut être déterminée en mesurant V_0 en choisissant $V_i \neq 0$ et les autres $V_j = 0$. Peu

importe le nombre d'observations effectuées, et le choix de V_i utilisé, seules les conductances relatives a_i pourront être déterminées, la valeur absolue des conductances γ_i ne sont pas mesurables de cette manière[46].

$$V_0 = \frac{\sum_{i=1}^{n} \gamma_i V_i}{\sum_{i=1}^{n} \gamma_i} = \sum_{i=1}^{n} a_i V_i \text{ avec } a_i = \frac{\gamma_i}{\sum_{k=1}^{n} \gamma_k} \tag{7.4}$$

Les a_i ainsi définies par (7.4) respectent $\sum_{i=1}^{n} a_i = 1$ et correspondent donc aux coordonnées barycentriques de V_0 dans le repère affine formé par les V_i. Pour pouvoir déterminer l'état de la mémoire par l'observation de V_0, il est nécessaire que les lieux de V_0 correspondants à chaque état soient disjoints. Dans la suite, nous nous restreignons au cas où chaque élément mémoire OxRAM peut être dans l'un des deux états, LRS et HRS. De plus, nous considérons que la conductance d'un élément mémoire OxRAM en HRS appartient à l'intervalle $[0, h]$ et celle d'un élément en LRS à l'intervalle $[l, 1]$. Fixer ainsi le maximum de conductance à 1 est sans perte de généralité puisque la valeur absolue des conductances est sans intérêt ici.

Nous montrons en Annexe C que tous les états sauf l'état où tous les éléments mémoires OxRAM sont en HRS sont distinguables si et seulement si $h < l^2$. L'état où tous les éléments mémoires OxRAM sont en HRS est quant à lui confondable avec n'importe quel autre état, il est donc nécessaire d'interdire cet état pour cette méthode de lecture. Cette méthode de lecture permet donc de distinguer $2^n - 1$ états sous la condition $h < l^2$ avec n le nombre d'éléments mémoires OxRAM par cellule mémoire. Cette condition implique que les états HRS et LRS soient suffisamment séparés par une zone de conductance interdite. Cette zone est caractérisée par $l - h$, or $l - h > l(1 - l)$. Donc, à valeur de conductance LRS minimale (l) constante, l'augmentation de la conductance maximale en LRS conduit

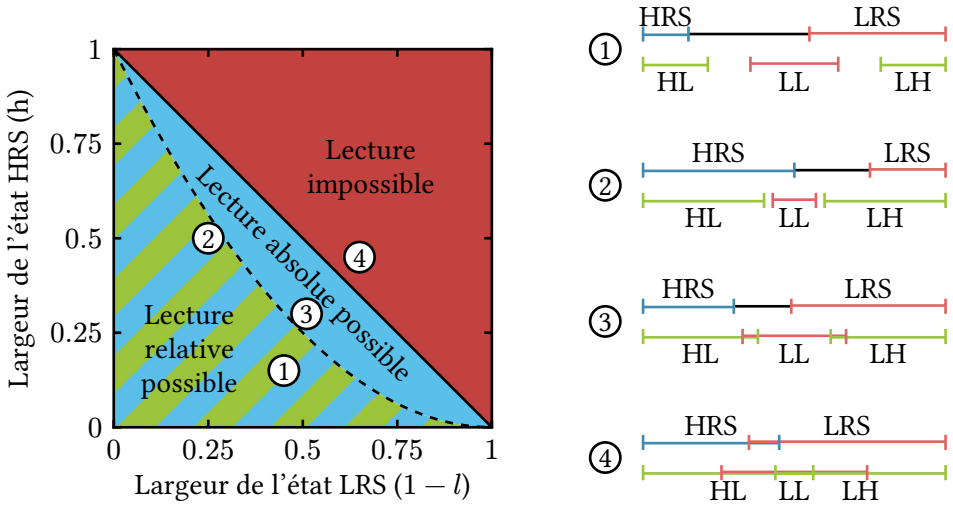


Figure 7.6 – À gauche, une représentation graphique de la contrainte $h < l^2$ en fonction des largeurs des intervalles de conductances associés aux états LRS et HRS. À droite, représentation des intervalles des lectures pour quatre exemples de choix de l et h particuliers avec, en haut, les états vus par une lecture absolue, en bas, les états vus par une lecture relative. Dans les exemples 1 et 2, la lecture relative est possible. Dans l'exemple 3 seule la lecture absolue est possible. Dans l'exemple 4 les états HRS et LRS ne sont pas séparés et ne permettent aucune lecture.

[46]Les valeurs absolues des conductances γ_i ne peuvent pas être déterminées par simple observation de V_0. En effet, considérons deux situations, la première avec les conductances γ_i et la seconde avec les conductances $k\gamma_i$. Dans ces deux situations, les valeurs prises par V_0 en fonction des autres potentiels V_i seront identiques. Ainsi il est impossible de déterminer le facteur k, et seul le rapport relatif entre les γ_i nous est accessible.

à une augmentation de la taille de la zone de conductance interdite, et donc à la diminution de la conductance HRS admissible.[47] Il est possible de respecter cette condition en pratique. La Figure 7.6 représente l'ensemble des combinaisons de choix de h (largeur de l'état HRS en terme de conductance normalisée) et de $1-l$ (largeur de l'état LRS en terme de conductance normalisée).

Il est possible de réduire la condition de séparabilité à $h < l$, c'est-à-dire que les états soient simplement disjoints en terme de conductance absolue, cependant seulement des ensembles de n états sont alors distinguables. La densité d'informations est alors fortement réduite. La lecture dite différentielle correspond à ce cas particulier avec $n = 2$. L'un des intérêts de la lecture différentielle est son effet sur le taux d'erreur [39]. En effet, imaginons que la distribution des HRS suit une loi normale $\mathcal{N}\left(\mu_{\text{HRS}}, \sigma^2_{\text{HRS}}\right)$ et celle des LRS une loi normale $\mathcal{N}\left(\mu_{\text{LRS}}, \sigma^2_{\text{LRS}}\right)$. En lecture simple, en notant τ le seuil de lecture, le taux d'erreur lorsque le dispositif est en HRS est la probabilité $\mathcal{E}_S = P(X > \tau) = \Phi\left(\frac{\mu_{\text{HRS}} - \tau}{\sigma_{\text{HRS}}}\right)$ avec $\Phi(x)$ la fonction de répartition d'une loi normale centrée réduite. Dans le cas de la lecture différentielle, en notant X_1 l'état du premier élément mémoire OxRAM supposé en HRS et X_2 l'état du second supposé en LRS, le taux d'erreur est la probabilité $\mathcal{E}_D = P(X_1 > X_2) = P(X_1 - X_2)$. En supposant X_1 et X_2 indépendants on a $X_1 - X_2 \sim \mathcal{N}\left(\mu_{\text{HRS}} - \mu_{\text{LRS}}, \sigma^2_{\text{HRS}} + \sigma^2_{\text{LRS}}\right)$ et donc $\mathcal{E}_D = \Phi\left(\frac{\mu_{\text{HRS}} - \mu_{\text{LRS}} - \tau}{\sqrt{\sigma^2_{\text{HRS}} + \sigma^2_{\text{LRS}}}}\right)$. En supposant $\sigma_{\text{HRS}} = \sigma_{\text{LRS}} = \sigma$ et en choisissant sans perte de généralité $\tau = 0$, $\mu_{\text{HRS}} = -\mu$ et $\mu_{\text{LRS}} = \mu$, on obtient $\mathcal{E}_S = \Phi\left(-\frac{\mu}{\sigma}\right)$ contre $\mathcal{E}_D = \Phi\left(-\sqrt{2}\frac{\mu}{\sigma}\right)$. Ainsi, l'effet sur le taux d'erreur de la lecture différentielle est de réduire la variance efficace, d'un facteur $\sqrt{2}$ dans ce cas. Pour les cas $n > 2$, on devrait pouvoir s'attendre à un effet similaire qui pourrait justifier la densité sacrifiée. Cependant, cet effet n'a pas d'avantage étudié dans cette thèse.

Intéressons-nous maintenant à comment nous pouvons concrètement effectuer la lecture et déterminer l'état de la cellule dans le cas où $h < l^2$. L'opération à la base de la lecture est l'application des tensions V_i et la mesure de la tension V_0 résultante. Cette opération nous apporte une information scalaire, qui correspond à une projection du point V_0 sur un axe particulier dans le repère (a_i). Cet axe est déterminé par le choix des tensions V_i appliquées. De la même manière que sur la Figure 7.4, cette projection peut être vue géométriquement comme la projection plaçant chacun des sommets V_i sur une droite tels que leur distance entre eux soit égale à leur différence de potentiel. En notant $v_i = \frac{V_i}{V}$ avec V une tension, l'axe sur lequel la projection se produit est celui normal à l'hyperplan défini par l'équation $\sum_i v_i a_i = c$ avec c une constante. Par comparaison de la mesure de V_0 avec cV, on peut déduire de quel côté de l'hyperplan se trouve le point V_0. En effectuant un ensemble de mesure en choisissant bien ces hyperplans, on peut déterminer si les coordonnées a_i du point V_0 sont dans un ensemble convexe arbitraire, et donc en particulier s'ils sont dans un certain lieu de V_0 correspondant à un état de la cellule. Pour effectuer les choix des V_i à appliquer, on peut s'intéresser aux hyperplans de séparation optimale déterminés dans la preuve de la condition de séparabilité en Annexe C. L'idée est d'utiliser ces hyperplans pour déterminer l'état de la cellule par une méthode de bissection. Un hyperplan est choisi comme point de départ, puis en fonction du résultat, l'hyperplan suivant est choisi de sorte à maximiser la quantité d'informations obtenue par la seconde mesure. Puis un troisième hyperplan est sélectionné de manière avantageuse et ainsi de suite jusqu'à ce qu'il ne reste plus qu'un seul état possible.

De façon intéressante, il semblerait que le sous-ensemble des hyperplans passant par les sommets V_i ou croisant les arêtes du simplex en $\frac{l}{1+l}$ suffisent à effectuer cette méthode de bijection en un nombre minimal d'étape. Pouvoir se limiter à cette famille d'hyperplans présente l'intérêt de simplifier l'implémentation de cette méthode. En effet, les choix de tension v_i associés à cette famille

[47] On constate ici que la largeur de la distribution LRS impacte directement les performances de cette méthode de lecture. Cela donne un exemple de l'intérêt de réduire celle-ci par la diminution de la conductance maximale en LRS comme nous l'avons fait au Chapitre IV.

d'hyperplans est l'ensemble $\left\{0, \frac{l}{1+l}, 1\right\}^n$ et la constante associée à ces hyperplans est toujours $c = \frac{l}{1+l}$. Cela signifie, que seulement trois sources de tension différentes seraient nécessaires : la masse, une source de tension référence, ici V, et une source de tension complémentaire également utilisée comme valeur de comparaison pour V_0 de valeur $\frac{l}{l+1}V$.

Cette partie s'est concentrée sur le cas où les éléments mémoires OxRAM ne peuvent être que dans deux états. Pour augmenter la densité d'informations, il peut être intéressant d'augmenter ce nombre d'états possible. Il est donc intéressant d'étudier l'applicabilité de cette méthode de lecture à ces cas. Prenons **le cas à trois états que l'on notera H, M et L** et avec 2 éléments mémoires OxRAM par cellule. La cellule a donc 9 états (HH, HM, MH, HL, LH, MM, ML, LM, LL). En appliquant cette méthode de lecture, **seul 5 de ces 9 états sont distinguables**. En effet, l'état HH est, comme précédemment superposé à tous les autres états. L'état HL est inclus dans l'état HM et ils sont donc indistinguables, de même pour LH et MH. Enfin, les états LL et MM sont également confondus, avec l'un inclus dans l'autre. Le sens de l'inclusion dépend de la largeur relative de l'intervalle de conductance octroyée aux états M et L. Ainsi, 4 états doivent être interdits pour permettre une lecture non ambigüe. De plus, pour pouvoir distinguer les états ML et LM de l'état LL, il est nécessaire que les intervalles de conductance soient suffisamment espacés et respectent des conditions[48] plus strictes que la condition $h < l^2$ du cas à deux états. Si ces conditions ne sont pas respectées, seuls 3 ou 4 états (en fonction des conditions non respectées) ne sont distinguables. Dans le cas multi-niveaux, où les éléments mémoires OxRAM sont programmés sur plus de 2 niveaux, cette méthode de lecture ne permet donc pas un stockage d'informations dense puisque presque la moitié des états possibles ne sont pas utilisables. Cependant, dans le cas correspondant à une lecture différentielle, où aucune condition sur les conductances n'est respectée, où seuls 3 états sont distinguables, elle permet encore de réduire le taux d'erreur de la lecture, en échange d'une réduction de la densité d'informations.

VII.2.5 Choix du nombre n d'éléments mémoires OxRAM par sélecteur

L'intérêt de n'utiliser qu'un transistor pour N éléments mémoires OxRAM est de pouvoir augmenter la taille du transistor sans réduire la densité mémoire, ou au contraire, augmenter la densité lorsque le transistor est le facteur limitant. Le transistor peut être limitant à la fois par sa taille minimale limitée par la technologie utilisée, mais surtout par son courant de saturation. En effet, la densité de courant passant dans le transistor est limitée. Comme les éléments mémoires OxRAM nécessitent un certain courant pour être écrits et effacés, il est nécessaire de dimensionner le transistor suffisamment grand. Cela définit une taille minimale pour le transistor. A partir de cette taille, on peut obtenir une estimation du nombre N d'éléments mémoires OxRAM devant partager un transistor pour pouvoir atteindre une densité donnée.

Cependant, la présence des autres éléments mémoires OxRAM sur le même transistor rajoute des contraintes. En effet, une partie du courant traversant le transistor correspond à des courants dits parasites, qui ne traversent pas l'élément mémoire OxRAM ciblé par l'écriture. Il est donc nécessaire de surdimensionner le transistor, ce qui peut conduire à une augmentation de la précédente estimation de N.

De façon générale, on peut rapprocher la densité de courant, grandeur linéique intrinsèque à la technologie des transistors[49], à une densité surfacique en divisant par la longueur du transistor (en prenant en compte la longueur du canal et la longueur des zones de diffusion des drains et sources). Ainsi, on peut obtenir la quantité de courant qu'un transistor peut conduire par unité de surface, que l'on peut normaliser par nombre d'élément mémoire OxRAM. Bien sûr, il est nécessaire de multiplier

[48] En notant l'intervalle de l'état H $[0, h]$, celui de l'état M $[m, M]$ et celui de l'état L $[l, 1]$, les conditions à respecter pour pouvoir distinguer les états ML et LM de LL ou MM sont $M < l^2$ et $M^2 < lm$. De plus, pour pouvoir distinguer les états ML et LM de HM et MH, il est nécessaire d'avoir $h < m^2$.

[49] On considère la longueur du canal minimale et la tension de grille maximale.

cette valeur par une estimation du taux d'utilisation effectif de l'actif, représentant la part de la surface active effectivement utilisée par le transistor, excluant la surface perdue à cause des diverses contraintes liées à la géométrie des transistors et à leur environnement, les contacts des grilles, l'espacement entre transistor et autres règles de dessin. Si cette quantité de courant est inférieure à la quantité de courant parasite ajouté par un élément mémoire OxRAM supplémentaire, il n'y a pas de solution, et les estimations successives de N divergeraient. Il serait alors nécessaire de diminuer la densité mémoire ciblée. On constate donc que la quantité de courant parasite induite par un élément mémoire OxRAM combinée à la densité de courant de la technologie de transistor utilisée impose une limite sur la densité mémoire envisageable. On peut agir sur cette limite en choisissant une technologie de transistor ayant une meilleure densité de courant, cependant ce choix est limité par la tension de claquage, en maximisant le taux d'utilisation effectif de l'actif, ou en essayant de réduire le courant parasite dû à un élément mémoire OxRAM en utilisant une stratégie de programmation essayant de le minimiser.

Par rapport à cette estimation, il est important de prévoir une marge. Plus la densité visée est proche de sa limite, plus le nombre N croît rapidement. Cependant, il est nécessaire de conserver N faible. En effet, N influence la quantité totale de courant parasite, et contraint donc les circuits périphériques devant fournir le courant. De plus, plus le nombre N est important, moins le potentiel du nœud central de la cellule est contrôlé par la SL, et plus le schéma d'écriture sera contraint par l'incertitude en résultant. Notons que ceci est vrai même si l'on considère un sélecteur proportionnellement plus grand.[50]

VII.3 Mise en œuvre de l'architecture 1TnR : Circuit *OX1TnR*

VII.3.1 Objectifs

Du fait du caractère aléatoire et des difficultés de modélisation des OxRAM, en particulier de leur dynamique d'écriture et d'effacement, il est nécessaire de faire une réalisation pratique pour pouvoir juger de l'intérêt de l'idée. Cela permettra également d'identifier d'éventuelles considérations pratiques importantes oubliées dans l'étude théorique.

Pour la réalisation de ce circuit, nous avons saisi une opportunité en utilisant une surface disponible sur une technologie 22 nm (mais 40 nm de taille minimale pour les premiers niveaux de métal). La surface est assez grande pour permettre l'incorporation de nombreuses matrices aux caractéristiques variables. Nous souhaitons donc réaliser un panaché de matrices couvrant un ensemble de variables d'ajustement. En particulier, nous souhaitons avoir des matrices ayant un nombre variable d'éléments mémoires OxRAM par cellule, différents types de sélecteurs et de tailles différentes. Comme nos ressources sont limitées, nous optons pour la réalisation d'une unique matrice dite objectif composée de cellules contenant 8 éléments mémoires OxRAM, le maximum envisagé, et utilisant une paire de transistors NMOS/PMOS comme sélecteur. Les autres éléments du panaché seront dérivés de cette matrice par une édition minimale, même si cela revient à avoir une densité plus faible ou à avoir une architecture clairement sous-optimale. L'intérêt du panaché est de pouvoir observer jusqu'où l'on peut aller, et quels sont les gains envisageables pour savoir si cette technique est à approfondir ou à laisser de côté dans le futur. La densité réelle des matrices est donc un intérêt secondaire ici. Dans la même idée, seule la conception circuit est considérée ici.

[50] On peut noter $\gamma_0 = N\frac{\gamma_A}{D}$ la conductance du sélecteur avec N le nombre de dispositifs par sélecteur, γ_A la conductance par unité de surface du sélecteur et D la densité ciblée. On note la conductance parasite $\gamma_p = (N-1)\gamma_0$ avec γ_0 la conductance d'un élément mémoire OxRAM (on peut prendre la plus grande envisagée pour avoir le pire cas). On a alors le rapport de conductance du sélecteur par rapport à celle parasite $\frac{\gamma_0}{\gamma_p} = \frac{\gamma_A}{D\gamma_0}\frac{N}{N-1}$, qui tend vers $\lim_{N\to\infty}\frac{\gamma_0}{\gamma_p} = \frac{\gamma_A}{D\gamma_0}$ alors que sa valeur pour $N = 2$ est $\frac{\gamma_0}{\gamma_p} = 2\frac{\gamma_A}{D\gamma_0}$. On a donc bien une réduction du contrôle du potentiel du nœud central par la SL.

VII.3.2 Contraintes technologiques

La technologie utilisée est une technologie hybride constituée d'un front-end SOI CMOS 40 nm commerciale ainsi que ses premiers 4 niveaux d'interconnexion métal et d'une technologie OxRAM du CEA-Leti déposé par-dessus entre le métal M4 et M5. Les contraintes les plus importantes à prendre en compte pour la réalisation d'une matrice dense dans cette technologie sont :

1. **Orientation des grilles fixes**

   L'orientation des grilles est fixée soit horizontale soit verticale pour l'intégralité du circuit (ou du moins pour une surface bien supérieure à l'une de nos matrices). Ainsi l'orientation des sélecteurs doit être identique dans toute notre matrice. Nous la considérerons verticale dans la suite.

2. **Continuité des zones actives**

   Pour améliorer la mobilité des trous (qui limite la conductivité des PMOS par rapport au NMOS), le substrat est mécaniquement contraint. Pour éviter la relaxation de ces contraintes, et la diminution des performances des transistors, les grilles des transistors utiles doivent être gardées à une certaine distance de la fin de leur zone active. A l'échelle de nos sélecteurs, cela représente environ deux longueurs de transistor (longueur du canal plus largeur du drain/source). Cela signifie que nos bandes d'actif horizontales (car nécessairement perpendiculaire aux grilles) ne doivent pas être interrompues dans la matrice sous peine de devoir environ quintupler la longueur de nos transistors. Cela signifie que l'on ne peut pas mélanger les NMOS et les PMOS sur une même bande d'actif.

3. **Transistor GO2**

   Du fait des tensions nécessaires pour le fonctionnement des OxRAM, les transistors GO2 (*Gate Oxide 2*), signifiant essentiellement oxyde de grille épais, sont utilisés car ils supportent des tensions maximales plus grandes par rapport à leur homologue GO1 (*Gate Oxide 1*). Ces transistors ont vocation à servir d'interfaces pour le numérique, ou plus généralement pour des circuits analogiques ne profitant pas de l'amélioration de la densité apportée par les transistors GO1. Ils sont donc moins denses, à la fois par nécessité d'avoir une tenue en tension plus élevée, et donc d'avoir une longueur de canal plus grande, mais également par une offre moins grande d'optimisation. En particulier, alors que les transistors GO1 proposent une structure particulière permettant de séparer électriquement une zone active en deux sans relaxer les contraintes mécaniques, les transistors GO2 ne le permettent pas. La seule option pour conserver la densité est alors soit de s'assurer de ne pas avoir besoin d'une telle séparation électrique, soit d'utiliser un transistor toujours ouvert comme séparateur et de subir les courants de fuite en résultant.

4. **Espacement entre NMOS et PMOS**

   Dans une technologie CMOS dite « bulk », le canal des transistors est le substrat lui-même. Les NMOS et les PMOS nécessitent un substrat de dopage opposé. Généralement le substrat de départ convient aux NMOS et les PMOS sont placés dans des puits où le substrat a été re-dopé localement. Aucun transistor ne peut être placé proche du périmètre de ces puits. L'espace minimal entre un PMOS et un NMOS est donc bien supérieur à celui entre deux NMOS ou deux PMOS. Cependant, la technologie utilisée ici est FDSOI (« Fully Depleted Silicon On Insulator »). Contrairement aux technologies bulk, le substrat d'une technologie FDSOI est composé de trois couches, une couche fine de silicium sur une couche d'isolant (oxyde de silicium) elle-même reposant sur un substrat silicium. Les transistors sont gravés sur la couche fine. L'isolant restreint la profondeur du canal des transistors, le gardant proche de la grille en surface pour que cette dernière ait un meilleur contrôle sur celui-ci. Le substrat de silicium a un intérêt mécanique, mais il permet également de polariser le canal au travers de l'oxyde épais et agit en quelque sorte comme une grille secondaire. Dans une technologie FDSOI, les transistors NMOS et PMOS peuvent être placés dans un même puits. Il est donc possible de les placer aussi proches l'un de l'autre que deux NMOS ou deux PMOS. L'inconvénient est alors que les deux transistors auront leurs grilles secondaires polarisées de la même manière. Comme l'effet de la polarisation est opposé entre un NMOS et un PMOS, l'un des transistors aura sa tension de seuil affaiblie pendant que l'autre l'aura renforcée. La tension de seuil

joue sur le compromis entre courant de fuite et courant de saturation, ou d'un autre point de vue, entre capacité à bloquer le courant ou à le conduire. En pratique, cela a pour conséquence de rendre les problématiques de courant de fuite plus importantes. Néanmoins, cela nous permet de mélanger au plus proche des NMOS avec des PMOS, seul moyen de rendre l'utilisation d'un sélecteur composé d'un NMOS et d'un PMOS possible sans utiliser une surface prohibitive.

VII.3.3 Conception des masques d'une matrice 1T8R

Comme point de départ, nous nous donnons comme objectif de réaliser notre matrice objectif avec la densité mémoires maximales permise par la technologie OxRAM utilisée. Pour rester en accord avec tous les schémas présentés jusqu'ici, nous considérerons les BL et GL verticales et les SL horizontales.

Nous commençons par fixer le placement des éléments mémoires OxRAM. Bien que le réseau hexagonal permet en théorie d'atteindre de plus grandes densités grâce à sa meilleure répartition des points dans l'espace, nous utiliserons un réseau carré. En effet, les règles de proximité de la technologie utilisée sont définies indépendamment sur les axes horizontal et vertical. Sous cette condition, un réseau hexagonal permet exactement la même densité qu'un réseau carré. Nous choisissons ce dernier car il permet des interconnexions plus simples.

Nous continuons avec les connections électriques directes des éléments mémoires OxRAM. Comme les BL ne sont connectées qu'à la périphérie de la matrice et à un des terminaux des éléments mémoires OxRAM, il est possible de les placer au-dessus des éléments mémoires OxRAM. Nous connectons donc chaque colonne d'élément mémoire OxRAM à une ligne métallique au-dessus d'eux que l'on nomme BL. De part notre architecture, il est nécessaire de connecter l'autre terminal des éléments mémoires OxRAM en groupe de N éléments horizontaux avec des lignes que l'on appellera ligne X ou nœud X dans la suite. La Figure 7.7 montre une telle disposition. Les lignes BL doivent pouvoir conduire le courant d'écriture ainsi que les courants parasites, ce qui peut représenter un courant important. Les règles de dessins contraignent la densité de métal et donc la largeur des pistes BL. Si la conductivité de la piste n'est pas assez élevée, ou si elle ne respecte pas les contraintes d'électro-migration, il est possible d'utiliser le niveau de métal supérieur pour dédoubler les pistes.

Il est important que la matrice soit composée d'un motif se répétant de manière régulière. Cela permet de pouvoir ajuster sa taille en ajoutant ou retirant des occurrences, cela simplifie le dessin, et surtout, cela assure une certaine régularité dans l'environnement des dispositifs, et donc une meilleure homo-

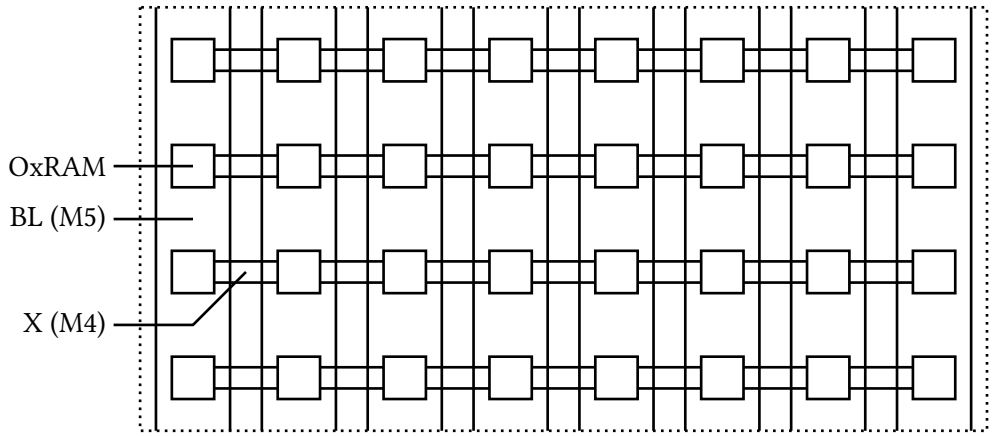


Figure 7.7 – Dessin des masques correspondants aux éléments mémoires OxRAM et aux niveaux métalliques les encadrant. Le trait en pointillé indique la limite du motif qui peut être répété par aboutement.

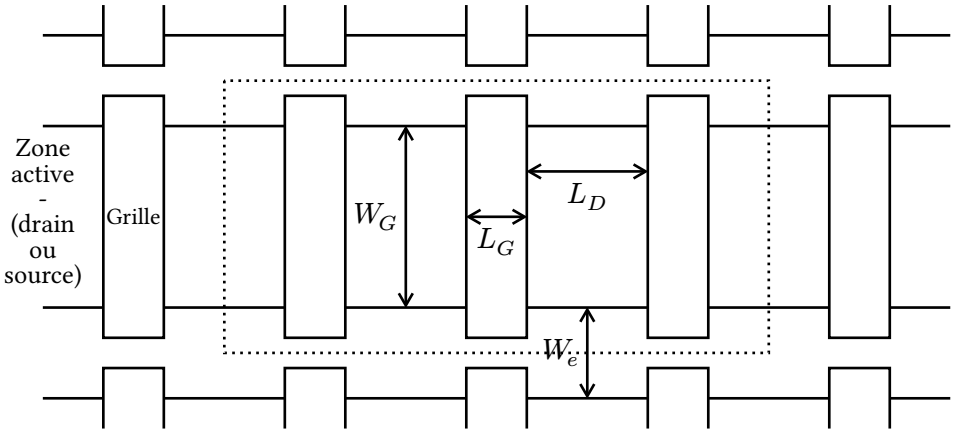


Figure 7.8 – Dessin d'un transistor à trois doigts au sein d'une matrice de transistors de taille proche du minimum et respectant la contrainte de continuité des zones actives. L_G est la longueur de la grille, L_D la longueur inter-grille, W_G la largeur de la grille et W_e la largeur entre deux bandes de transistors. Le rectangle en pointillé dénote la maille associée à un transistor dans ce réseau carré de transistors.

généité de leurs caractéristiques. Comme la disposition de nos éléments mémoires OxRAM est figée[51], le motif mesurera nécessairement un nombre entier d'éléments mémoires OxRAM de large et de long. Le plus simple serait que le motif corresponde à un groupe d'éléments mémoires OxRAM appartenant à la même cellule. Cependant, la géométrie allongée des groupes (8 éléments par 1) rend cette solution peu envisageable car elle induirait une faible efficacité d'utilisation de la surface disponible du point de vue des transistors. Pour mieux comprendre, nous nous penchons sur la géométrie des transistors.

La géométrie des transistors proches de leur taille minimale est à la fois contrainte par différentes règles de dessin permettant d'assurer qu'ils soient fabricables, mais également par notre volonté de maximiser la largeur effective des transistors sur la surface disponible. La Figure 7.8 représente un transistor MOS faisant partie d'une matrice de transistor MOS respectant la contrainte de continuité des zones actives précédemment évoquées ainsi que diverses grandeurs lui étant associées. La largeur de la grille W_G multipliée par le nombre de doigts (c'est-à-dire le nombre de grilles) du transistor est la grandeur que l'on cherche à maximiser. Nous allons considérer les directions horizontale (ou longueur dans la suite) et verticale (ou largeur dans la suite) de façon distincte car elle n'interagissent pas directement ensemble et présentent des contraintes différentes.

En commençant par la largeur, c'est-à-dire la direction verticale sur la Figure 7.8, la distance W_e entre deux zones actives a tout intérêt à être minimisée, car la surface non active ne participe pas à la conduction. La largeur de la grille W_G est la grandeur que l'on souhaite maximiser dans cette direction. Elle peut être ajustée de manière continue. Dans l'optique de maximiser W_G à l'échelle de la matrice (qui a une largeur fixée par le nombre de SL), il est nécessaire de réduire le nombre de fois où une largeur W_e doit être « gâchée » pour séparer deux bandes d'actif. Il est donc pertinent d'utiliser des bandes d'actif de plusieurs élément mémoire OxRAM de large.

Regardons maintenant l'autre direction, la longueur, c'est-à-dire la direction horizontale sur la Figure 7.8. La longueur de la grille L_G ainsi que la longueur inter-grille L_D ont tout intérêt à être minimisées. En effet, la longueur L_D ne participe en rien à la conduction du transistor, cependant elle limite le nombre de grilles que l'on peut disposer sur une surface donnée. La longueur de la grille L_G quant à elle, en plus de limiter le nombre de grilles, est inversement proportionnelle à la conductivité

[51] Du simple fait de notre objectif arbitraire de réaliser notre matrice objectif avec la densité mémoire maximale permise par la technologie OxRAM utilisée.

du transistor. Elle est donc fixée à sa valeur minimale qui est limitée par la tension de claquage. Nous pouvons ajuster ces longueurs pour mieux aligner le motif si nécessaire, mais toute divergence à la valeur minimale induira une réduction de la conductivité atteignable des transistors, et donc des sélecteurs. A l'échelle de la matrice, on a tout intérêt à ce que le motif se répétant ait une longueur au plus proche mais supérieure à un nombre entier de longueurs de transistor. Ainsi, la conductivité du transistor n'augmente pas de façon continue dans cette direction, mais de façon discrète.

En somme, la conductivité du transistor croît continûment avec la largeur disponible, mais de façon discrète selon la longueur disponible. L'efficacité de l'occupation de la surface croît, quant à elle, avec la largeur réservée à une bande de transistor et décroît avec la différence entre la longueur du motif et la longueur discrète associée. Pour maximiser la conductivité des sélecteurs, il faut maximiser l'efficacité de l'occupation.

Un motif correspondant aux groupes d'éléments mémoires OxRAM appartenant à une même cellule mesurerait un élément mémoire OxRAM de haut. Cette hauteur n'est pas assez grande pour n'accommoder ne serait-ce qu'une grille dans la direction de la longueur. La hauteur devrait donc correspondre à la largeur du transistor. Cependant dans ce cas, W_G serait inférieure à W_e et l'efficacité de l'occupation de la surface serait très faible car plus de la moitié de la largeur ne serait pas exploitée. Et cela, en ne considérant qu'un transistor par motif, alors que l'on souhaite en avoir deux, un NMOS et un PMOS. **Il est donc nécessaire de considérer un motif de plusieurs éléments mémoires OxRAM de haut**, et donc recouvrant plusieurs groupes d'éléments. Cela signifie également qu'une même bande d'actif devra servir pour plusieurs SL. Ce dernier point a des conséquences sur la compacité atteignable dans la direction de la longueur.

En effet, la longueur d'un transistor, ou plutôt celle occupée par une occurrence d'une grille n'est $L_G + L_D$ que si les zones de diffusion de part et d'autre de cette grille peuvent être partagées avec les grilles voisines. Si les grilles voisines font partie du même transistor, il suffit d'alterner les drains et les sources pour respecter cette contrainte. Cependant, à la transition entre deux transistors, ce n'est possible que si les deux transistors partagent leur drain ou leur source, comme sur la partie gauche de la Figure 7.9. Comme tous les drains des transistors sont séparés les uns des autres, ils ne peuvent pas être mis en commun. De plus, comme nous venons de le voir, plusieurs SL correspondent à une même bande d'actif. Par conséquence, au sein du motif, aucun transistor ne partage la même SL. Ainsi

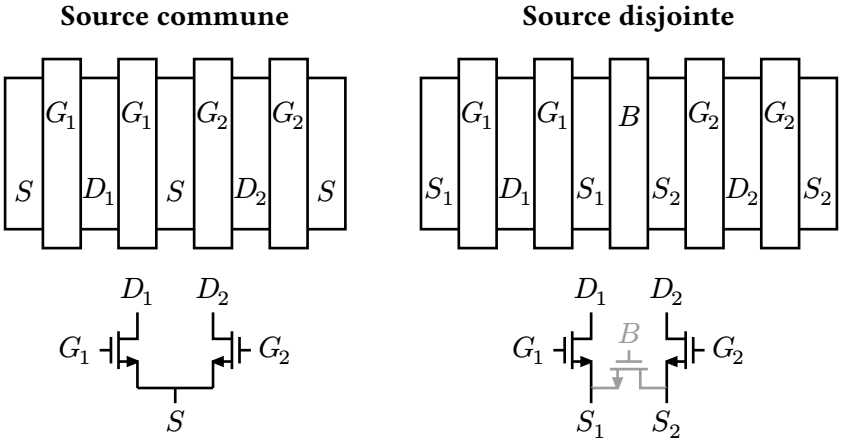


Figure 7.9 – Illustration de la problématique de dessin de deux transistors sur une même bande d'actif. A gauche, le cas où les deux transistors partagent leur source (cela pourrait également être leur drain). A droite, le cas où les deux transistors ne partagent ni leur source ni leur drain et requièrent donc une grille de séparation, noté B, polarisée pour ne pas laisser passer de courant entre les deux zones de diffusion correspondant à S_1 et S_2.

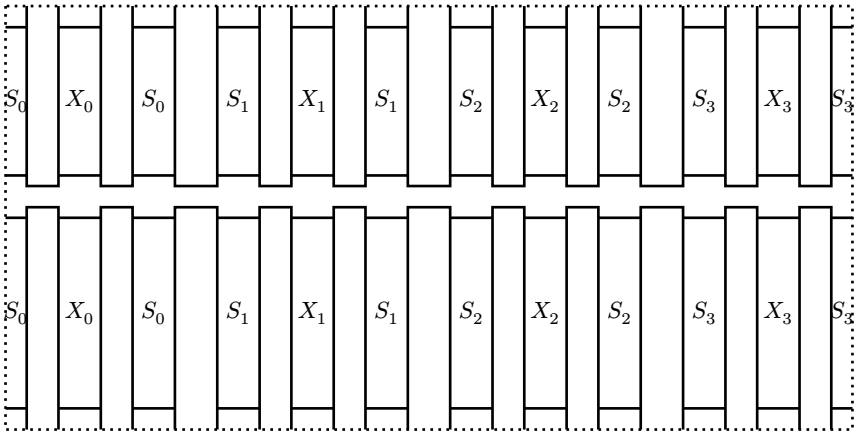


Figure 7.10 – Dessin simplifié des niveaux transistors. Seul l'actif et les grilles sont dessinés. Le texte indique à quoi chaque zone de diffusion est logiquement connectée, avec S_x la source numéro x et X_x le nœud interne de la cellule numéro x. Le trait en pointillé indique l'étendue du motif qui peut être répété par aboutement après réflexion.

ils ne peuvent pas non plus être combinés par la source. Comme on ne peut pas couper l'actif, le seul moyen est d'utiliser une grille fixée au bon potentiel pour bloquer la conduction entre deux zones actives adjacentes, comme illustré sur la partie droite de la Figure 7.9. Cela réduit alors le nombre de grilles disponibles pour la constitution de nos sélecteurs. En particulier, si n SL se partagent une même bande d'actif, au moins $n - 1$ grilles de séparation sont nécessaires. On a donc intérêt à limiter le nombre de SL se partageant une même bande d'actif, ce qui correspond à la convention que l'on a choisie précédemment, c'est-à-dire SL horizontale et grille verticale.

Avec ces informations, on peut maintenant simuler l'efficacité de différentes tailles de motif, en particulier selon la direction des longueurs. Avec la taille des éléments mémoires OxRAM utilisés ainsi que les dimensions propres à la technologie CMOS utilisée, un motif d'une longueur de 8 éléments et d'une largeur de 4 éléments présente une efficacité intéressante. C'est le motif que nous avons choisi pour notre matrice objectif. La longueur supplémentaire disponible est utilisée pour augmenter la longueur des grilles de séparation pour réduire leur courant de fuite. La Figure 7.10 présente le dessin des niveaux relatifs au transistor obtenue pour respecter cette taille de motif. Comme notre matrice objectif contient des sélecteurs constitués d'un NMOS et d'un PMOS connectés en parallèle, deux bandes d'actif sont présentes dans chaque motif.

Enfin, il ne reste plus qu'à faire les interconnections des drains des transistors au nœud X des groupes d'éléments mémoires OxRAM et de sortir les lignes GL et SL. Plus simple à écrire qu'à faire en réalité ! En effet, beaucoup de choses doivent se croiser et se connecter au front-end. Une chose à ne pas sous-estimer est la place que prend un via d'un niveau de métal vers le front-end sur tous les niveaux de métal intermédiaires. Du point de vue de l'encombrement, il est donc intéressant de distribuer les signaux ayant le plus de connexions sur les niveaux les plus bas. Les niveaux plus haut offrent quant à eux une conductivité plus élevée. Les GL et les signaux de polarisation des grilles de découpage de l'actif sont mis sur les deux premiers niveaux de métaux M1 et M2. Ces deux niveaux sont également utilisés pour connecter les paires de transistors NMOS et PMOS en parallèle. Les SL sont positionnées sur le troisième niveau M3. Le quatrième M4 est déjà utilisé par les lignes X connectées sous les éléments mémoires OxRAM.

La Figure 7.11 illustre la géométrie obtenue pour ces niveaux d'interconnexions. Les SL ne sont pas droites car elle doivent laisser de la place aux via connectant les lignes X aux drains qui sont verticalement contraints par le placement des lignes X. Les grilles des sélecteurs sont connectées au M1 par le bord du motif (entre deux transistors NMOS ou PMOS adjacents). Lorsque le sélecteur est

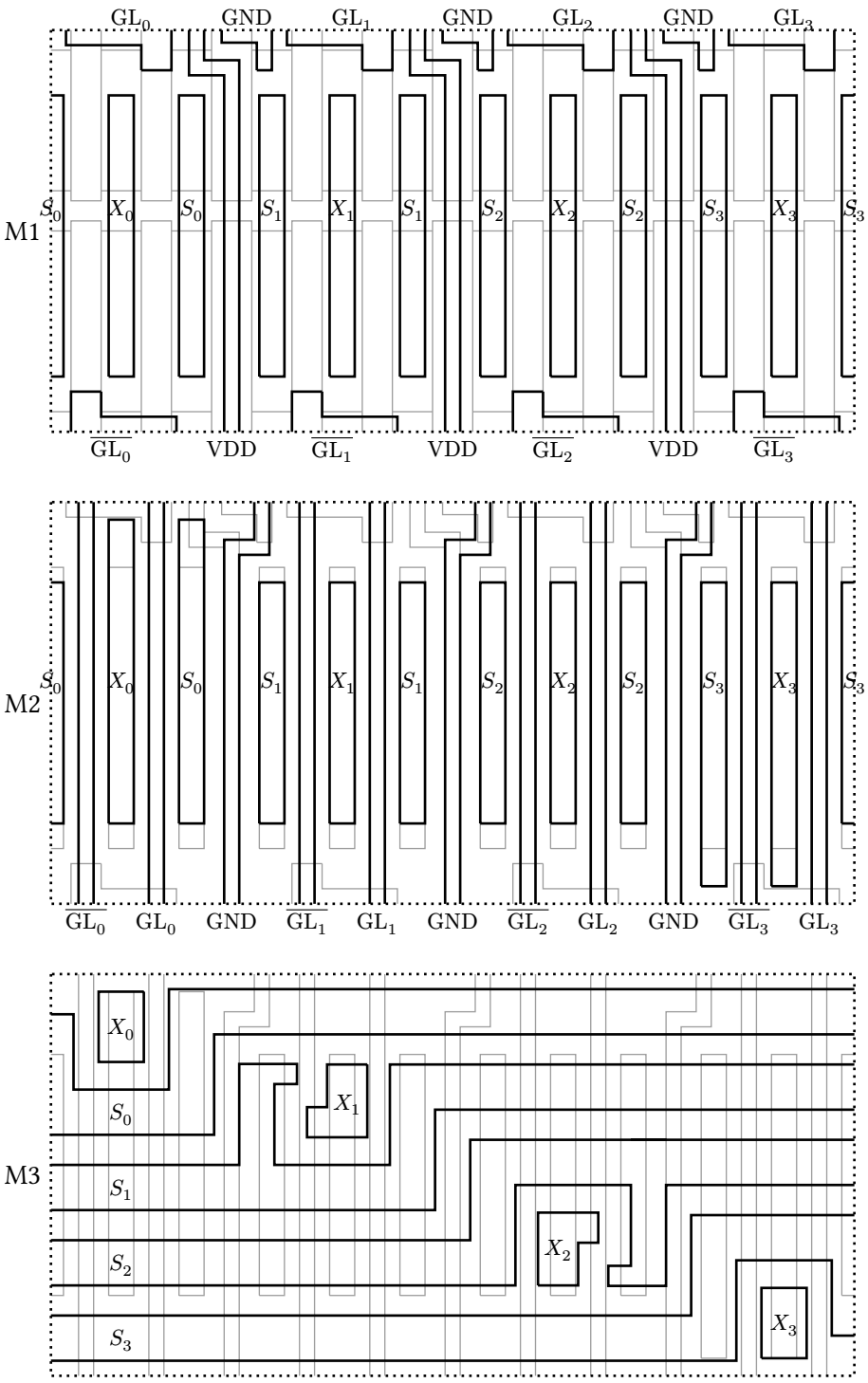


Figure 7.11 – Dessin des niveaux d'interconnexion entre le front-end (les grilles, drains et sources des sélecteurs) et le niveau métallique M4 connecté aux éléments mémoires OxRAM (représenté sur la Figure 7.7). Les trois niveaux d'interconnexion sont représentés séparément en trait noir. Les traits gris représentent le niveau en-dessous de celui dessiné. Le trait en pointillé indique l'étendue du motif qui peut être répétée par aboutement après réflexion.

composé d'un NMOS et d'un PMOS, deux lignes GL sont nécessaires par sélecteur, l'un étant l'opposé logique de l'autre. Elles sont toutes deux distribuées au niveau M2 en passant haut-dessus des grilles des sélecteurs. Deux tensions sont également nécessaires pour polariser les grilles de découpage de l'actif. Elles sont distribuées l'une par le niveau M1 et l'autre par le niveau M2 et passent toutes deux au-dessus des grilles qu'elles polarisent.

Ce motif est répétable sous forme de réseau carré à condition de le refléter pour faire correspondre chaque arête avec elle-même.

VII.3.4 Dérivation des autres matrices

A partir de la matrice objectif réalisée dans la partie précédente, nous dérivons de nombreuses autres architectures. L'objet de ces architectures n'est pas de repousser les limites de ce qu'il est possible de faire, mais plutôt de proposer à moindre coût un panel plus large de matrices pour l'expérimentation de cette architecture. Peut-être que 8 éléments mémoires OxRAM par cellule sont trop élevés ? Peut-être que les courants parasites dus à la présence de PMOS dans le sélecteur sont gênants pour l'estimation de certains paramètres ? Ou peut-être pouvons-nous pousser l'architecture encore plus loin et pouvons-nous nous satisfaire d'un sélecteur plus petit ?

Nous avons identifié 4 principaux paramètres que l'on souhaite faire varier :

1. **Nombre d'éléments mémoires OxRAM par cellule** (8R, 4R, 2R, 1R ou 0R)
   Pour varier le nombre d'éléments mémoires OxRAM par cellule, nous retirons un certain nombre de colonnes d'éléments mémoires OxRAM ainsi que leur BL du dessin. Nous ne modifions rien d'autre. La densité mémoire est donc proportionnellement réduite. La variante 0R correspond à une matrice dépourvue d'élément mémoire OxRAM. A la place elle contient des via spécialement positionnés pour permettre la mesure de la conductivité des lignes BL et SL. Cette disposition de via et son utilisation sont expliquées dans la sous-partie suivante.
2. **Nombre de SL** (128 SL, 64 SL, 32 SL, 16 SL)
   Comme nous l'avons vu dans la partie théorique, la quantité de courant parasite croît quasi-linéairement avec le nombre de SL. Pour permettre l'observation de leur évolution en fonction du nombre de SL, nous incluons 4 tailles de matrices. Pour ce faire, nous retirons autant de lignes d'éléments mémoires OxRAM que nécessaire sans rien modifier de plus.
3. **Type de sélecteur** (NMOS, N+PMOS asymétrique, N+PMOS symétrique, $\frac{1}{2}$N+PMOS symétrique)
   Nous avons identifié le besoin probable d'utiliser une porte de transmission (noté N+PMOS) en tant que sélecteur pour permettre le bon fonctionnement de notre architecture car la réduction de la conductance d'un transistor MOS seul lorsque proche de la masse ou de l'alimentation réduit le contrôle du nœud central des cellules mémoires. Cependant, nous souhaitons tout de même pouvoir expérimenter des matrices avec des sélecteurs constitués d'un unique NMOS. Pour effectuer ce changement, nous remplaçons le PMOS par un NMOS en conservant les deux bandes d'actif séparées, bien qu'elles pourraient être réunies. Proche de la masse, la conductivité du NMOS seul est plus grande que celle de la porte de transmission. Nous implémentons également trois types de porte de transmission. La première précédemment présentée octroie plus de surface au PMOS pour compenser la plus faible mobilité des trous face à celle des électrons. La seconde répartit équitablement la surface entre les deux transistors, améliorant ainsi la conductivité près de la masse au dépend de celle près de l'alimentation. Et la troisième fait de même mais n'occupe que la moitié de la surface disponible.
4. **Taille de la périphérie** (BL_{mux}, $\frac{1}{2}BL_{mux}$, SL_{mux}, $\frac{1}{2}SL_{mux}$)
   Nous n'en avons pas encore parlé, mais la périphérie de la matrice, en particulier les multiplexeurs reliant les BL et SL sélectionnées d'une part et non sélectionnées d'autre part en deux lignes électriques distinctes accessibles depuis l'extérieur du circuit présentent une conductivité limitée. La présence de courant important dans la matrice, en particulier dû aux courants parasites indique que la conductivité des multiplexeurs doit être élevée sous peine de perdre une partie importante de la tension dessus. Pour limiter ce risque et retirer le plus possible cet aspect de l'équation, la périphérie a été largement sur-dimensionnée, si bien qu'elle occupe une surface plus importante que le corps de la matrice elle-même. Pour nous donner au moins un second point d'observation selon cette dimension, nous avons également inclus des matrices dont la taille des multiplexeurs a été divisée par deux.

Tableau 7.3 – Caractéristiques des matrices disponibles dans le circuit *OX1TnR*

	barrette 0	barrette 1	barrette 2	barrette 3	barrette 4	barrette 5
matrice 0	1T0R	1T0R	1T0R	1T0R	1T0R	1T0R
	128 SL	128 SL	128 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS sym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	$\frac{1}{2}BL_{mux}$ SL_{mux}	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	BL_{mux} SL_{mux}
matrice 1	1T1R	1T1R	1T1R	1T1R	1T1R	1T1R
	128 SL	128 SL	128 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS sym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	$\frac{1}{2}BL_{mux}$ SL_{mux}	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	BL_{mux} SL_{mux}
matrice 2	1T8R	1T8R	1T8R	1T8R	1T8R	1T8R
	128 SL	128 SL	128 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS sym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	$\frac{1}{2}BL_{mux}$ SL_{mux}	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	BL_{mux} SL_{mux}
matrice 3	1T8R	1T8R	1T8R	1T8R	1T8R	1T8R
	64 SL	64 SL	64 SL	64 SL	64 SL	64 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS sym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	$\frac{1}{2}BL_{mux}$ SL_{mux}	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	BL_{mux} SL_{mux}
matrice 4	1T8R	1T8R	1T8R	1T4R	1T4R	1T4R
	32 SL	32 SL	32 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS sym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	$\frac{1}{2}BL_{mux}$ SL_{mux}	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	BL_{mux} SL_{mux}
matrice 5	1T8R	1T8R	1T8R	1T0R	1T0R	1T0R
	16 SL	16 SL	16 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS asym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$
matrice 6	1T4R	1T4R	1T4R	1T1R	1T1R	1T1R
	128 SL	128 SL	128 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS asym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$
matrice 7	1T4R	1T4R	1T4R	1T8R	1T8R	1T8R
	64 SL	64 SL	64 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS asym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$
matrice 8	1T4R	1T4R	1T4R	1T8R	1T8R	1T8R
	32 SL	32 SL	32 SL	64 SL	64 SL	64 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS asym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$
matrice 9	1T2R	1T2R	1T2R	1T4R	1T4R	1T4R
	128 SL	128 SL	128 SL	128 SL	128 SL	128 SL
	N+PMOS sym	NMOS	N+PMOS asym	N+PMOS asym	N+PMOS asym	$\frac{1}{2}$N+PMOS sym
	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} SL_{mux}	BL_{mux} $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$	$\frac{1}{2}BL_{mux}$ $\frac{1}{2}SL_{mux}$

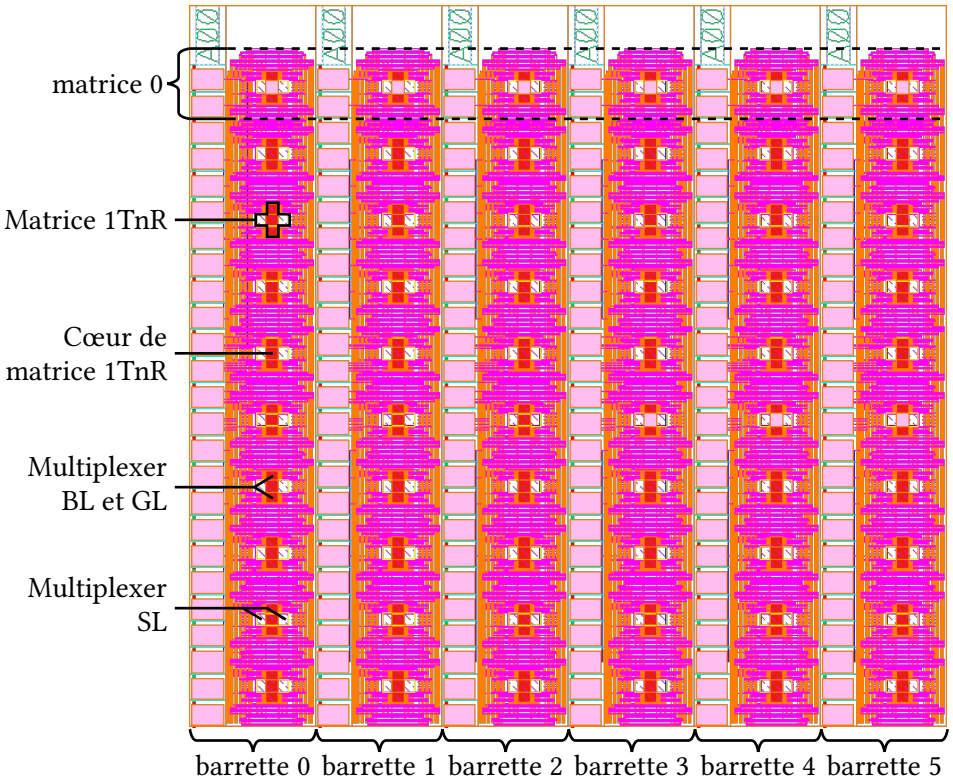


Figure 7.12 – Vue d'ensemble du dessin du circuit *OX1TnR*.

Le Tableau 7.3 présente l'ensemble des caractéristiques des matrices implémentées dans le circuit *OX1TnR*. La matrice objectif est entourée en rouge. La Figure 7.12 illustre le circuit *OX1TnR* dans son intégralité. Il est composé des 6 barrettes totalement indépendantes et contenant chacune 10 matrices.

VII.3.5 Motif pour la mesure des impédances des lignes

Du fait des courants parasites importants au sein de la matrice, l'impédance des lignes BL et SL, mais aussi celle de la périphérie, sont des grandeurs importantes à déterminer même pour une « petite » matrice d'au plus 16k éléments telle que la nôtre. Pour ce faire, nous incluons des matrices dépourvues d'éléments mémoires OxRAM (1T0R), ayant à la place des vias court-circuitant certaines lignes ensemble. Le motif de ces court-circuits est conçu pour permettre la détermination des impédances visibles au sein de la matrice ainsi que leur évolution en fonction de l'adresse. La Figure 7.13 illustre ces connexions.

Une particularité de la périphérie de notre matrice est qu'elle est répartie équitablement sur ses quatre côtés. Une BL sur deux est contrôlée par le haut, les autres par le bas, et de même pour les SL sur les côtés gauche et droit. Pour estimer l'impédance des multiplexeurs seuls, nous court-circuitons une paire de multiplexeurs adjacents ensemble sur chaque côté et nous nous assurons que leur ligne ne soit connectée à rien d'autre. La mesure s'effectue ensuite en sélectionnant une adresse associée à l'un des deux multiplexeurs et en mesurant l'impédance entre la ligne BL/SL sélectionnée et la ligne BL/SL correspondante aux non-sélectionnées. La valeur obtenue est alors le double de l'impédance d'un multiplexeur puisque deux sont présents en série sur le chemin de mesure. Une fois cette valeur obtenue, elle pourra être soustraite dans les prochaines mesures pour isoler la partie qui nous intéressera.

Les impédances des lignes BL et SL n'ont aucune raison d'être les mêmes. En effet, ces deux types de lignes n'ont pas la même géométrie et ne sont pas sur les même niveaux métalliques. La longueur effective (celle parcourue par le courant) des BL et SL dépend également de l'adresse. Pour nous

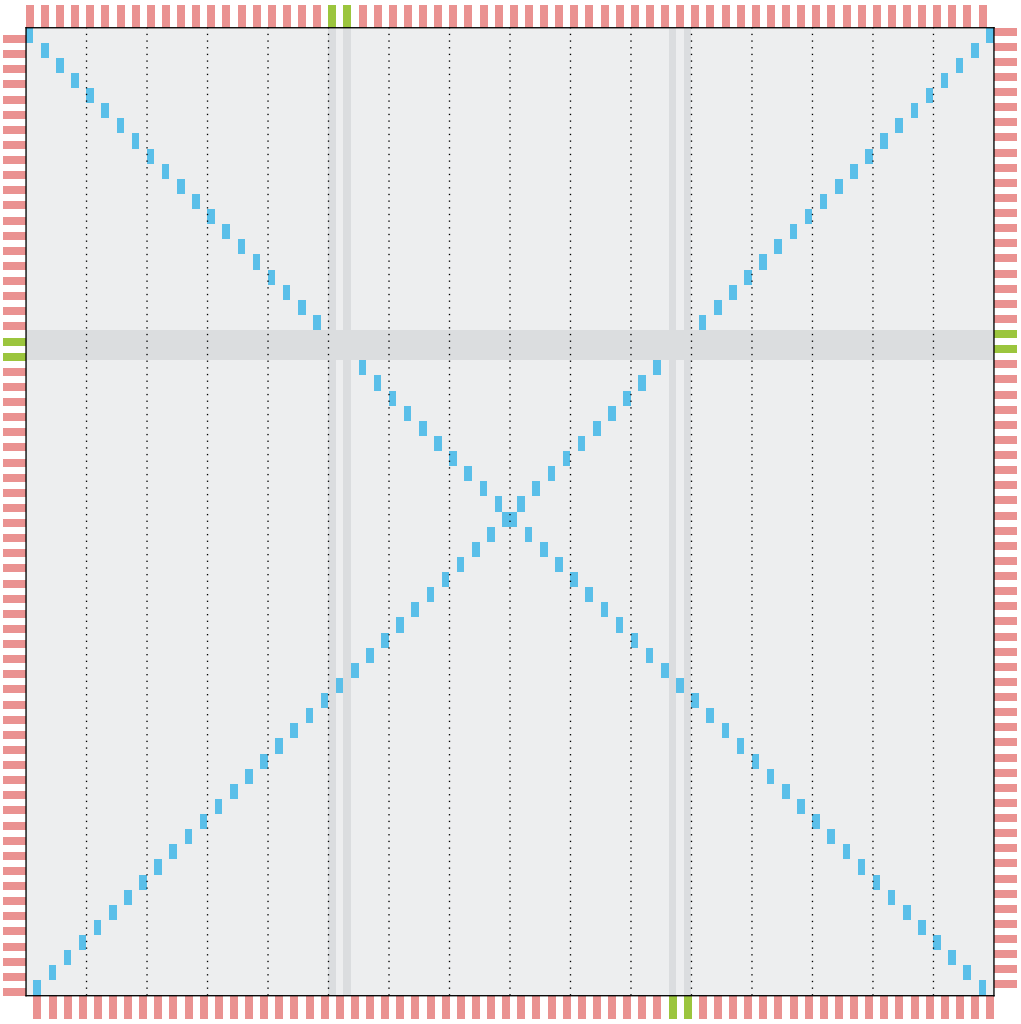


Figure 7.13 – Structure d'une matrice 1T0R. Chaque point représente une adresse. Les points bleus correspondent à un court-circuit en place d'un élément mémoire OxRAM. Ceux en gris indiquent l'absence de connexion. Les zones rouge pâle et verte en périphérie représentent les multiplexeurs d'accès aux BL en haut et en bas, et aux SL à gauche et à droite. Les multiplexeurs verts sont les paires court-circuitées entre elles. Les lignes et colonnes en gris foncé sont celles sur lesquelles les multiplexeurs d'accès sont court-circuités et donc celles sur lesquelles il ne doit y avoir aucune autre connexion. Enfin, les lignes en pointillé positionnées toutes les 8 colonnes séparent la matrice en 16 zones représentant chacune une colonne de cellule mémoire partageant une même GL.

permettre de séparer ces deux composantes, nous effectuons la mesure de l'impédance perçue entre la BL et la SL sur différents chemins. L'objectif est d'obtenir, d'un côté un ensemble de mesure selon des chemins de longueur BL + SL constante, et d'un autre selon des chemins de longueur BL - SL nulle.

Pour mesurer l'impédance d'un chemin, une des adresses court-circuitées (point bleu sur Figure 7.13) doit être sélectionnée. La tension sur les BL et SL non sélectionnées doivent être maintenues égales pour qu'aucun courant ne circule dans les autres court-circuits de la colonne de cellule. Un potentiel différent peut être appliqué sur la BL sélectionnée, et la SL sélectionnée peut être gardée en haute impédance. Le courant circule alors entre la BL sélectionnée, au travers du court-circuit voisin à celui adressé, jusqu'aux SL non sélectionnées. L'intérêt d'avoir des paires de court-circuits partageant la même BL, de garder la SL sélectionnée en haute impédance et d'effectuer la mesure entre la BL sélectionnée et les SL non sélectionnées (plutôt que la SL sélectionnée directement) et de pouvoir utiliser la

SL sélectionnée comme mesure Kelvin. En effet, comme aucun courant ne circule sur la SL sélectionnée, son potentiel sera égal au potentiel à l'endroit où elle rencontre le chemin du courant, indépendamment de l'impédance présente entre ce dernier et le point de mesure. Ainsi, cela nous permet de déterminer à la fois l'impédance totale sur le chemin du courant (multiplexeur BL + BL + sélecteur + SL + multiplexeur SL) mais également le potentiel en amont du sélecteur, c'est à dire la proportion entre la partie du chemin avant et après (ici multiplexeur BL + BL et sélecteur + SL + multiplexeur SL). Cela nous permet donc d'obtenir l'impédance de la BL seule en soustrayant l'impédance du multiplexeur précédemment établi.

Enfin, pour séparer l'impédance du sélecteur de celle de la SL, nous pourrons l'estimer à partir de l'ensemble des mesures aux différentes adresses en supposant qu'elle est proportionnelle à la distance.

Les courants de fuite entre SL dus aux transistors de découpage de l'actif peuvent être mesurés en appliquant une différence de potentiel entre une SL sélectionnée et les SL non sélectionnées. La BL sélectionnée doit alors être maintenue au potentiel de la SL sélectionnée et celui des BL non sélectionnées à celle des SL non sélectionnées.

A l'aide de ces matrice 1T0R, nous devrions pouvoir mesurer et séparer les impédances des multiplexeurs, des sélecteurs, des BL et des SL. Pour le BL et les SL nous pourrons également obtenir leur évolution au travers de la matrice.

VII.4 Conclusion

Bien que l'idée d'une architecture accommodant plus d'éléments mémoires OxRAM par sélecteur n'est pas nouvelle, nous avons proposé dans ce chapitre une forme particulière jamais observée dans la littérature. Nous l'avons identifiée et distinguée de l'architecture communément référencée en montrant ces avantages sur certains points et son équivalence sur d'autres. Ces deux principaux avantages sont le meilleur contrôle des nœuds internes aux cellules ainsi qu'un courant parasite indépendant d'une des dimensions de la matrice lors d'une opération parallèle sur plusieurs éléments mémoires OxRAM.

Nous avons ensuite décrit les différents aspects à considérer lors de l'écriture des éléments. Nous avons discuté des quatre axes d'optimisation possibles pour l'écriture et de leur compromis, particulièrement entre l'un d'entre eux, la tension maximale applicable sur l'élément à écrire, et les trois autres visant à minimiser différents types de courant parasite. Nous avons également illustré une méthode graphique permettant de prendre en compte l'effet de la conductivité finie du sélecteur sur les tensions perçues par les différents types d'éléments au court de l'écriture.

Après l'écriture, nous avons discuté de la lecture. Nous avons proposé deux méthodes de lecture très différentes. La première permettant d'accéder à la conductance précise de chaque élément est semblable à la méthode utilisée dans les chapitres précédents. La seconde permettant une lecture relative, sans référence absolue de courant ou de conductance. Elle correspond à la généralisation de la lecture différentielle à un nombre plus élevé d'éléments. Cependant, dans l'optique de maximiser la densité, nous avons également montré que cette méthode permettait l'identification de $2^n - 1$ états, avec n le nombre d'éléments mémoires OxRAM par cellule, sous une condition nécessaire et suffisante concernant les supports des distributions de la conductance des états LRS et HRS.

Enfin, nous avons passé en revue la conception d'une implémentation de cette idée de matrice 1TnR. Notre description s'est concentrée sur la méthode de conception utilisée pour la réalisation des dessins des masques du cœur de la matrice. Nous avons également évoqué l'ensemble des versions de matrice disponibles permettant un test plus large et quelque peu paramétrique.

Des résultats expérimentaux n'ont pas pu être obtenus dans le cadre de cette thèse par un manque de temps. Nous espérons que le circuit *OX1TnR* pourra effectivement être utilisé pour approfondir l'idée de l'architecture 1TnR.

Bibliographie

[39] M. Bocquet *et al.*, « In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks », in *2 018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 20. doi: 10.1109/IEDM.2018.8614639.

[94] J. Zhang *et al.*, « A 3D RRAM using stackable 1TXR memory cell for high density application », in *2 009 International Conference on Communications, Circuits and Systems*, juill. 2009, p. 917-920. doi: 10.1109/ICCCAS.2009.5250369.

[95] J. Yang *et al.*, « A 28 nm 1,5 Mb Embedded 1T2R RRAM with 14.8 Mb/mm2 using Sneaking Current Suppression and Compensation Techniques », in *2 020 IEEE Symposium on VLSI Circuits*, juin 2020, p. 1-2. doi: 10.1109/VLSICircuits18222.2020.9163035.

[96] C.-W. S. Yeh et S. S. Wong, « Compact One-Transistor-N-RRAM Array Architecture for Advanced CMOS Technology », *IEEE Journal of Solid-State Circuits*, vol. 50, nº 5, p. 1299-1309, mai 2015, doi: 10.1109/JSSC.2015.2402217.

[97] L. Zhao, L. Jiang, Y. Zhang, N. Xiao, et J. Yang, « Constructing fast and energy efficient 1TnR based ReRAM crossbar memory », in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, mars 2017, p. 58-64. doi: 10.1109/ISQED.2017.7918293.

[98] E. R. Hsieh *et al.*, « Four-Bits-Per-Memory One-Transistor-and-Eight-Resistive-Random-Access-Memory (1T8R) Array », *IEEE Electron Device Letters*, vol. 42, nº 3, p. 335-338, mars 2021, doi: 10.1109/LED.2021.3055017.

[99] E. Hsieh *et al.*, « High-Density Multiple Bits-per-Cell 1T4R RRAM Array with Gradual SET/RESET and its Effectiveness for Deep Learning », in *2 019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 35. doi: 10.1109/IEDM19573.2019.8993514.

Conclusion

Chapitre I – Au-delà de leur fonction de stockage, les mémoires sont au cœur de toute machine de calcul. Dans un monde exigeant une capacité de calcul croissante, les mémoires sont devenues l'élément limitant. Bien que les technologies actuelles continuent de progresser, elles ne permettent pas de suivre les progrès réalisés par les éléments de calcul.

La recherche de nouveaux composants mémoires offrant de meilleures performances ou permettant la création de nouvelles architectures est une des voies explorées pour rompre avec cette trajectoire. C'est dans ce contexte que s'inscrit la recherche sur les mémoires résistives dont l'OxRAM fait partie.

Leur fonctionnement semblable à un memristor a inspiré de nombreuses nouvelles architectures de calcul. Mais ce sont plutôt les architectures multiplicateur-accumulateur exploitant leur non-volatilité ainsi que leur fonctionnement en tant que résistance variable qui semblent prometteuses. En effet, ces opérations sont omniprésentes dans de nombreuses applications de calcul, dont celle de la très en vogue intelligence artificielle.

Ces travaux de thèse ont exploré deux principaux axes d'amélioration des performances des mémoires OxRAM au niveau du circuit les intégrant.

Le contrôle de niveaux de résistivité des éléments mémoires OxRAM lors de l'écriture semble critique et motive la première partie de cette thèse concernant l'écriture des OxRAM (Chapitres II à V). Cette partie traite de l'amélioration de l'uniformité de l'écriture au travers de l'utilisation de sources électriques de natures différentes.

L'applicabilité de ces solutions repose également sur la densité d'intégration accessible aux éléments mémoires OxRAM et motive donc la seconde partie de cette thèse (Chapitres VI et VII). Cette partie explore l'augmentation de la densité d'intégration des éléments mémoires OxRAM à l'aide d'une nouvelle architecture de matrice multiplexant l'utilisation d'un sélecteur pour plusieurs éléments mémoires OxRAM.

Chapitre II – Dans ce chapitre d'introduction sur la thématique de l'écriture des éléments mémoires OxRAM, nous identifions le courant limite d'écriture en SET comme un paramètre central du niveau de conductance atteint en LRS. Nous passons ensuite en revue les différentes techniques d'écriture référencées par la littérature sous le prisme de différents critères de performance préalablement définis. Nous en concluons que les méthodes d'écriture présentes dans la littérature sont presque toutes basées sur l'utilisation de sources de tension comme excitation principale. Elles se distinguent cependant par

l'utilisation d'une variété d'algorithmes de programmation. Ce manque de données sur l'utilisation d'autres sources d'excitation électrique est à l'origine de l'orientation de nos travaux. Comme le courant d'écriture représente un paramètre important, nous nous intéressons à l'écriture à l'aide d'une source de courant. Puis, nous nous tournons vers l'écriture à l'aide d'une capacité chargée.

Chapitre III – Nous commençons l'étude de l'écriture en courant par une approche théorique. Nous notons la relation d'équivalence entre écriture en tension limitée en courant et écriture en courant limitée en tension dans un contexte statique. La différence entre ces deux écritures réside dans les effets dynamiques. L'effet au cœur de l'écriture en courant est le pic de courant dû à la décharge capacitive lors de la transition de HRS à LRS. Pour débuter l'étude expérimentale de l'écriture en courant, nous exploitons un circuit réalisé lors d'une précédente thèse. Bien qu'il n'ait pas été conçu pour, son architecture permet de réaliser nos premières expériences. Comme la capacité présente entre la source et l'élément mémoire OxRAM joue un rôle important dans l'écriture, il était important de pouvoir utiliser une source co-intégrée aux éléments mémoires OxRAM.

L'utilisation d'un circuit en-dehors de son cas d'application envisagé nous a imposé certaines limites. En particulier, nous avons eu des difficultés à réaliser l'étape de Forming de façon répétable. Le nombre limité d'éléments mémoires OxRAM disponibles nous a également poussé à ne considérer que la statistique cycle-à-cycle et non dispositif-à-dispositif. Nous étudions expérimentalement l'influence de l'amplitude de l'impulsion en courant utilisé, puis celle de sa durée en réalisant des séries d'écriture à différentes amplitudes et durées. Nous en déduisons que le pic de courant lors de l'écriture participe à l'écriture et permet d'atteindre des conductances plus importantes que celles obtenues avec une écriture en tension limitée à un courant de même amplitude que l'impulsion de courant utilisée. D'un autre côté, la durée de l'impulsion semble sans effet, du moins sur la gamme de temps observée. Ces observations correspondent à l'analyse théorique initialement conduite.

Nous nous intéressons ensuite à l'endurance des éléments mémoires OxRAM écrits en courant pour vérifier que le pic de courant ne causerait pas un vieillissement accéléré. Nous avons pu mettre en évidence qu'aucun signe de vieillissement prématuré n'est apparu lors de 200 000 cycles. Comparativement à une méthode d'écriture en tension sur la même technologie d'élément mémoire OxRAM, mais sur un circuit différent, la variance et la médiane du niveau de conductance atteint en LRS par notre méthode sont plus grandes. La plage de conductance atteinte est similaire du côté des conductances faibles, mais plus large du côté des conductances fortes. Ainsi, cette variance plus forte ne réduit pas directement la fenêtre de lecture, on obtient un taux d'erreur similaire entre l'écriture en tension et en courant.

Pour finir, nous simulons *a posteriori* l'effet qu'un algorithme de vérification aurait eu sur notre écriture en courant. La large variance avec une médiane décalée vers les hautes conductances engendrées par notre méthode améliore l'efficacité de l'algorithme de vérification. Cela permet d'obtenir des marges de lecture plus importantes que l'écriture en tension pour un même taux de rejet.

Ces résultats doivent être mis en perspective avec les limitations auxquelles nous avons fait face. En particulier, ils ont été obtenus sur un nombre très réduit d'éléments mémoires OxRAM. Leur répétabilité n'a donc pas été vérifiée. De plus, le choix des éléments mémoires OxRAM utilisés a potentiellement été biaisé par les difficultés

rencontrées pour effectuer l'opération de forming. En effet, ils sont ceux que l'on a réussi à former et sur lesquels l'écriture semblait fonctionner.

Chapitre IV – Pour poursuivre l'étude de l'écriture en courant, et lever les limitations rencontrées par notre première expérience, nous réalisons un circuit spécifiquement conçu pour l'expérimentation de l'écriture en courant. Ce circuit offre une plus large gamme de sources de courant, un nombre plus important d'éléments mémoires OxRAM et supporte également l'écriture en tension. Cela permettra l'étude des statistiques dispositif-à-dispositif et la comparaison *in situ* entre des méthodes d'écriture en courant et d'écriture en tension. Nous développons également sa plateforme de test qui assure un suivi précis des opérations effectuées sur chaque élément mémoire OxRAM et qui permet l'expérimentation sans instrument de laboratoire. Nous décrivons son fonctionnement et détaillons l'implémentation des algorithmes d'écriture utilisés dans la suite.

Sur le plan expérimental, nous commençons par explorer l'opération de forming aussi bien en tension qu'en courant, avec et sans limitation du courant. Directement après le forming, les distributions observées dépendent fortement de la méthode utilisée. Cependant, après un millier de cycles d'écriture, ces différences s'estompent. Une particularité persistante est cependant observée sur la méthode de forming en courant sans limitation de courant. Deux populations d'éléments mémoires OxRAM se distinguent, l'une opérant à une conductance moyenne plus haute que l'autre. Cette expérience nous a également permis de sélectionner notre méthode de forming pour la suite de nos expériences.

Nous nous intéressons ensuite à l'écriture en SET. Nous comparons des méthodes d'écriture en tension et en courant, soit progressives, soit simple impulsion. La méthode d'écriture en courant simple impulsion présente la même caractéristique que le forming en courant sans limitation de courant, c'est-à-dire une séparation des éléments mémoires OxRAM en deux populations se comportant de manière identique, mais avec deux moyennes de conductance différentes. Par ailleurs, les méthodes d'écriture en courant progressives présentent des distributions LRS moins étendues que leurs homologues en tension. Nous concluons avec une comparaison quantitative de la proportion d'éléments mémoires OxRAM dépassant certains taux d'erreur sur 1 000 cycles. L'ensemble des méthodes d'écriture considérées présente un taux d'éléments défectueux (+ 10 % d'échec de l'écriture) similaire, comme si ces éléments étaient dysfonctionnels par construction, peu importe la méthode d'écriture utilisée. En excluant ceux-ci, l'écriture progressive vérifiée en courant ainsi qu'en tension présentant une conductance LRS moyenne la plus forte montrent le moins d'erreurs d'écriture, avec un léger avantage pour celle en courant. Cela est dû à une plus large fenêtre de lecture car il reste plus d'espace entre la distribution LRS et HRS. A seuil de lecture comparable, la conclusion est similaire et la différence avec les autres méthodes reste significative.

Chapitre V – Après l'écriture en courant, nous nous intéressons à une écriture utilisant une capacité préchargée. L'intérêt d'une telle écriture serait de bien contrôler la quantité d'énergie dissipée sur l'élément mémoire OxRAM lors de son écriture. La taille de la capacité nécessaire est une question cruciale pour juger de l'intérêt de cette méthode. En effet, la quantité de capacité est directement proportionnelle à la surface nécessaire, elle influe donc sur la densité et le coût de cette solution. Des travaux précédents avaient mis en évidence qu'une telle écriture était possible avec une capacité de 330 pF. Cependant, cette démonstration utilisait des capacités externes à la puce intégrant les éléments

mémoires OxRAM. Une partie de l'énergie était donc perdue par effet de partage de charge dans les interconnexions. Nous pouvons donc penser qu'une solution entièrement intégrée devrait pouvoir fonctionner avec une capacité plus petite.

Nous présentons un circuit développé dans l'objectif de répondre à cette question et d'expérimenter avec cette écriture. Il est conçu pour permettre l'écriture en tension, utilisée comme base de comparaison, et différentes sortes d'écriture capacitive. En particulier, nous présentons une méthode d'écriture capacitive dite avec recharge dans l'objectif de réduire encore plus la capacité nécessaire. Nous prévoyons également d'essayer d'utiliser la capacité intrinsèque de la matrice mémoire comme tout ou partie de la capacité d'écriture. Le circuit a pu être finalisé, simulé et envoyé en fabrication pendant cette thèse, mais il n'a pu être ni testé ni exploité expérimentalement. Nous ne présentons donc pas de résultats expérimentaux et nous ne pouvons pas apporter de réponse aux questions ayant motivé ce développement.

Ceci conclut la première partie de cette thèse qui aura montré que des écritures différentes sont envisageables. En particulier, l'écriture en courant présente des résultats comparables à l'écriture en tension, voire peut être meilleurs. Elle pourrait permettre des implémentations plus simples à l'échelle du système, du moins l'option est maintenant ouverte. L'écriture capacitive reste en suspens, mais nous avons fait un pas en direction d'une réponse sur sa faisabilité. Ces travaux sont à poursuivre.

Chapitre VI – Dans ce chapitre d'introduction à la problématique de l'intégration dense des éléments mémoires OxRAM, nous constatons que la réduction de la taille des éléments mémoires OxRAM augmente la tension de forming et influence peu leur courant d'écriture ou leur conductivité, aussi bien en LRS qu'en HRS. Ce second point est au cœur du problème de l'augmentation de la densité d'intégration des éléments mémoires OxRAM car il conduit à une augmentation de la densité de courant nécessaire, ce qui représente un obstacle à la réduction de la taille des composants nécessaires autour des éléments mémoires OxRAM.

Une seconde problématique de l'intégration dense est celle du multiplexage des accès. Le multiplexage est essentiel pour permettre l'accès à des milliers ou millions d'éléments mémoires OxRAM avec un nombre réduit de connexions électriques. Dans le cas de dipôles tels nos éléments mémoires OxRAM, nous montrons en quoi l'organisation sous forme de matrice carrée est optimale. Comme les éléments mémoires OxRAM sont conducteurs, elle introduit cependant un problème de courant de fuite.

Nous présentons les trois solutions généralement adoptées face à ce problème. La première consiste à l'atténuer sans vraiment le résoudre définitivement par une taille de matrice limitée et des schémas de programmation et de lecture particuliers. La deuxième ajoute un élément passif fortement non-linéaire en série avec l'élément mémoire OxRAM. Enfin, la troisième utilise un dispositif actif, en série avec l'élément mémoire OxRAM, dont la conductance peut être modifiée à l'aide d'une troisième terminale, comme un transistor MOSFET par exemple. Ces trois méthodes présentent chacune leur inconvénient. La première limite la taille de matrice réalisable, la deuxième requiert des dispositifs encore peu matures et la troisième est limitée en densité par la taille du transistor. En pratique, c'est cette troisième solution qui est la plus répandue et adoptée. Cependant, elle impose une limite à la densité d'intégration des éléments mémoires OxRAM de par la densité de courant limitée des transistors.

Chapitre VII – Une quatrième solution a déjà été proposée dans la littérature. Elle est hybride entre

la première, consistant à faire avec les courants de fuite, et la troisième utilisant des sélecteurs actifs, généralement des transistors MOSFET. L'idée est d'utiliser des transistors MOSFET pour améliorer la sélectivité, mais d'avoir plusieurs éléments mémoires OxRAM par transistor pour augmenter la densité lorsqu'il est l'élément limitant. Dès lors que l'on franchit ce pas, la présence de courant de fuite est inévitable et nous devons donc exploiter des méthodes issues de la première solution, d'où le caractère hybride de cette quatrième solution.

Nous débutons par une revue des trois organisations de matrice envisageables. Nous montrons la relative équivalence entre deux d'entre elles, et discutons des similitudes et différences avec la troisième. Nous montrons que cette troisième option est supérieure aux deux autres car elle permet de réduire la quantité de courant de fuite en fonctionnement. C'est donc cette solution que nous retenons pour la réalisation de notre circuit d'étude.

Nous décrivons ensuite la problématique de l'écriture d'un élément mémoire OxRAM particulier au sein de cette matrice. Nous montrons que l'on peut effectuer différents choix permettant d'optimiser dans une direction parmi quatre, à savoir : maximiser la tension d'écriture, minimiser le courant dans le transistor de l'élément mémoire OxRAM sélectionné, minimiser le courant de fuite dans la ligne de l'élément sélectionné et minimiser le courant parasite dans les autres lignes. Nous discutons du rôle joué par la conductance finie du transistor et de son effet sur l'écriture d'abord de façon analytique, puis nous illustrons un outil géométrique facilitant cette analyse.

Nous nous intéressons ensuite à la lecture. Nous proposons deux méthodes, l'une permettant d'accéder à l'information complète de la conductance de chaque élément mémoire OxRAM, la seconde ne permettant d'accéder qu'à l'état binaire de chaque élément. La première consiste à lire la conductance en mesurant le courant circulant dans le transistor de l'élément mémoire OxRAM sélectionné, lorsqu'une tension connue est appliquée à ses bornes. La présence des autres éléments mémoires OxRAM sur ce même transistor perturbe cette observation d'autant plus que la conductance du transistor est proche de la somme de celle des éléments mémoires OxRAM. Nous détaillons comment l'information peut être déduite d'un ensemble de n mesures avec n le nombre d'éléments mémoires OxRAM par transistor. La seconde méthode consiste à observer la tension aux bornes du transistor lorsqu'aucun courant ne le traverse. Elle ne permet pas d'obtenir la valeur absolue des conductances mais uniquement leur valeur relative entre celles des différents éléments mémoires OxRAM. Nous montrons que cette deuxième méthode permet de distinguer $2^n - 1$ des 2^n états possibles d'un groupe de n éléments mémoires OxRAM sous une condition limitant la proximité des deux états de conductance possibles. Nous montrons que cette condition est indépendante du nombre n. Nous discutons également de la version de cette méthode étendue à des éléments mémoires OxRAM multi-niveaux qui s'avère être très limitée et essentiellement inapplicable.

Enfin, nous discutons du choix du nombre d'éléments mémoires OxRAM par transistor. Il doit être gardé le plus bas possible, tout en réalisant la densité souhaitée avec la taille de transistor nécessaire pour conduire le courant requis.

Nous présentons ensuite le circuit réalisé pour expérimenter cette architecture de matrice d'éléments mémoires OxRAM. Nous nous concentrons principalement sur la description du dessin du cœur de la matrice. Nous expliquons les contraintes physiques

> et géométriques que nous devons respecter. Nous montrons comment elles mènent à devoir séparer la disposition des transistors et celle de leurs éléments mémoires OxRAM en deux découpages de l'espace différents. Le motif unitaire constituant la matrice par réflexion et répétition contient donc plusieurs transistors ainsi que plusieurs groupes d'éléments mémoires OxRAM. Nous montrons comment chaque niveau de métal est organisé pour permettre le routage des différents signaux appartenant aux quatre cellules mémoires partageant le motif unitaire. Nous présentons ensuite les variations de matrice que nous avons implémentées pour permettre l'observation de l'effet de divers paramètres de la matrice tels que le nombre d'éléments mémoires OxRAM par cellule, le nombre de cellules par groupe partageant un même réseau de courants de fuite ou encore le type de transistor utilisé. Nous détaillons également un motif permettant la mesure des impédances des lignes métalliques constituant la matrice. Ce circuit ayant fini sa fabrication proche de la fin de cette thèse, l'expérimentation des méthodes d'écriture et de lecture décrites n'a pas pu être réalisée. Après mon départ du CEA-Leti, quelques mesures préliminaires ont été effectuées pour montrer que les fonctionnalités de base du circuit étaient opérationnelles. Cependant, l'écriture des éléments mémoires OxRAM n'a pas été expérimentée à ma connaissance.

En conclusion, durant cette thèse, deux grandes thématiques autour des OxRAM ont été étudiées du point de vue du circuit électronique intégrant ces dispositifs. Trois circuits intégrés ont été conçus pour permettre respectivement l'étude de l'écriture en courant, celle de l'écriture capacitive et l'expérimentation d'une nouvelle architecture de matrice 1TnR mettant l'accent sur la densité.

L'étude de l'écriture en courant a montré le fort potentiel de cette méthode. L'étude comparative entre écritures en tension et écritures en courant devrait être poursuivie en s'intéressant à d'autres qualités d'une écriture d'élément mémoire OxRAM telles que : la rétention, la sensibilité à la relaxation, l'effet sur l'endurance et la stabilité des niveaux de conductance des états HRS et LRS au cours de la vie de l'élément mémoire OxRAM. Les qualités d'une écriture ne se limitent pas à leur effet sur l'élément mémoire OxRAM, mais aussi aux contraintes qu'elles imposent au système les intégrant. Il serait intéressant de comparer l'intégration d'une solution d'écriture en courant face à l'intégration d'une solution d'écriture en tension, en particulier sur les points tels que : la vitesse d'écriture, la consommation par écriture, le délai d'allumage des circuits d'écriture ou leur consommation en *stand-by*, la surface occupée pour les circuits d'écriture en particulier en fonction du nombre d'écritures parallèles effectuable, le niveau de précision ou de calibration requis par le circuit d'écriture pour atteindre les performances d'écriture souhaitées et la sensibilité du circuit d'écriture aux parasites variables dus au placement de l'élément mémoire OxRAM à écrire au sein de la matrice mémoire. L'écriture en courant en RESET pourrait également être étudiée. Les résultats préliminaires que nous avons montrent une tendance à sur-effacer les éléments mémoires OxRAM, réduisant drastiquement leur durée de vie. Un meilleur contrôle de la tension limite pourrait être un premier élément à étudier dans ce contexte.

L'architecture de matrice 1TnR présentée dans la seconde partie semble *a priori* tournée vers les applications de stockage de par sa densité accrue. Néanmoins, elle est en théorie également adaptée aux méthodes de calcul dans la mémoire telles que la multiplication-accumulation analogique. Sa densité profiterait en effet à ces méthodes qui excellent lorsque l'intégralité des poids de multiplication peut être stockée dans la mémoire OxRAM. En effet, cela élimine le besoin d'écrire les éléments mémoires OxRAM pendant le calcul. Dans ce contexte, la difficulté est de bien contrôler l'écriture pour pouvoir atteindre le niveau de précision de conductance requis. L'inconvénient de l'architecture 1TnR est la complexité de ce contrôle, à cause des chemins de conduction parasites. Cependant, dans le cas où l'écriture n'est à effectuer qu'une seule fois, l'utilisation d'une procédure d'écriture plus complexe est envisageable. On pourrait imaginer une procédure essayant de converger petit à petit vers l'état

souhaité. En somme, on pourrait dire que l'architecture 1TnR déplace le compromis entre la simplicité d'écriture et la densité du côté de la densité. Il serait donc intéressant d'étudier cet aspect dans un cas d'application pratique. Toutes les propriétés généralement d'intérêt pour les éléments mémoires OxRAM, telles que la rétention, l'endurance, la relaxation, sont potentiellement affectées par cette architecture 1TnR et représentent donc des points d'études à approfondir. Dans cette optique de densité, le développement des sélecteurs back-end permettant d'implémenter des architectures denses, sans front-end et sans courant de fuite est à surveiller. En effet, l'architecture 1TnR présentée devrait être dépassée sans difficulté par des architectures intégrant avec succès ces sélecteurs back-end. En effet, ces architectures résolvent la même problématique, mais de deux manières différentes. D'un côté l'approche 1TnR profite de la maturité des technologies CMOS en dépit des contraintes qu'elles imposent. De l'autre côté, l'approche avec des sélecteurs back-end cherche à développer un composant sur mesure réglant le problème de sélectivité et souffre pour l'instant d'un manque de maturité qui attend d'être résolu. De ce point de vue, l'architecture 1TnR est une solution temporaire.

Bibliographie

[1] C. Lam, « AMD's Zen 4, Part 2: Memory Subsystem and Conclusion ». [En ligne]. Disponible sur: https://chipsandcheese.com/p/amds-zen-4-part-2-memory-subsystem-and-conclusion?utm_source=publication-search

[2] I. Jo, « Toward Ultra-Low Latency SSDs: Analyzing the Impact on Data-Intensive Workloads », *Electronics*, vol. 13, nº 1, 2024, doi: 10.3390/electronics13010174.

[3] McCalpin et D. John, « Memory bandwidth and System Balance in HPC Systems », in *International Conference for High Performance Computing, Networking, Storage, and Analysis*, 2016.

[4] P. Kocher *et al.*, « Spectre Attacks: Exploiting Speculative Execution ». [En ligne]. Disponible sur: https://arxiv.org/abs/1801.01203

[5] M. Lipp *et al.*, « Meltdown ». [En ligne]. Disponible sur: https://arxiv.org/abs/1801.01207

[6] B. Gras, K. Razavi, H. Bos, et C. Giuffrida, « Translation Leak-aside Buffer: Defeating Cache Side-channel Protections with TLB Attacks », in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, MD: USENIX Association, août 2018, p. 955-972. [En ligne]. Disponible sur: https://www.usenix.org/conference/usenixsecurity18/presentation/gras

[7] Y. Yarom, D. Genkin, et N. Heninger, « CacheBleed: a timing attack on OpenSSL constant-time RSA », *Journal of Cryptographic Engineering*, vol. 7, p. , 2017, doi: 10.1007/s13389-017-0152-y.

[8] G. Dessouky *et al.*, « HardFails: Insights into Software-Exploitable Hardware Bugs », in *28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA: USENIX Association, août 2019, p. 213-230. [En ligne]. Disponible sur: https://www.usenix.org/conference/usenixsecurity19/presentation/dessouky

[9] P. Jain *et al.*, « 13.2 A 3,6 Mb 10,1 Mb/mm2 Embedded Non-Volatile ReRAM Macro in 22 nm FinFET Technology with Adaptive Forming/Set/Reset Schemes Yielding Down to 0,5 V with Sensing Time of 5 ns at 0,7 V », in *2 019 IEEE International Solid- State Circuits Conference - (ISSCC)*, févr. 2019, p. 212-214. doi: 10.1109/ISSCC.2019.8662393.

[10] J. Yang *et al.*, « 24.2 A 14 nm-FinFET 1 Mb Embedded 1T1R RRAM with a 0.022µm2 Cell Size Using Self-Adaptive Delayed Termination and Multi-Cell Reference », in *2 021 IEEE International Solid- State Circuits Conference (ISSCC)*, févr. 2021, p. 336-338. doi: 10.1109/ISSCC42613.2021.9365945.

[11] C.-C. Chou *et al.*, « An N40 256K×44 embedded RRAM macro with SL-precharge SA and low-voltage current limiter to improve read and write performance », in *2 018 IEEE International Solid - State Circuits Conference - (ISSCC)*, févr. 2018, p. 478-480. doi: 10.1109/ISSCC.2018.8310392.

[12] B. Giraud *et al.*, « Benefits of Design Assist Techniques on Performances and Reliability of a RRAM Macro », in *2 023 IEEE International Memory Workshop (IMW)*, mai 2023, p. 1-4. doi: 10.1109/IMW56887.2023.10145984.

[13] C. Ho *et al.*, « Integrated HfO2-RRAM to achieve highly reliable, greener, faster, cost-effective, and scaled devices », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 2. doi: 10.1109/IEDM.2017.8268314.

[14] Y. Liu *et al.*, « 4.7 A 65 nm ReRAM-enabled nonvolatile processor with 6× reduction in restore time and 4× higher clock frequency using adaptive data retention and self-write-termination nonvolatile logic », in *2 016 IEEE International Solid-State Circuits Conference (ISSCC)*, janv. 2016, p. 84-86. doi: 10.1109/ISSCC.2016.7417918.

[15] S. W. Keckler, W. J. Dally, B. Khailany, M. Garland, et D. Glasco, « GPUs and the Future of Parallel Computing », *IEEE Micro*, vol. 31, nº 5, p. 7-17, sept. 2011, doi: 10.1109/MM.2011.89.

[16] G. Sassine, C. Cagli, J.-F. Nodin, G. Molas, et E. Nowak, « Novel Computing Method for Short Programming Time and Low Energy Consumption in HfO2 Based RRAM Arrays », *IEEE Journal of the Electron Devices Society*, vol. 6, p. 696-702, 2018, doi: 10.1109/JEDS.2018.2830999.

[17] C.-C. Wang, R. G. B. Sangalang, I.-T. Tseng, Y.-J. Chiu, Y.-C. Lin, et O. L. J. A. Jose, « A 1.0 fJ energy/bit single-ended 1 kb 6T SRAM implemented using 40 nm CMOS process », *IET Circuits, Devices & Systems*, vol. 17, nº 2, p. 75-87, 2023, doi: 10.1049/cds2.12141.

[18] L. Chua, « Memristor-The missing circuit element », *IEEE Transactions on Circuit Theory*, vol. 18, nº 5, p. 507-519, sept. 1971, doi: 10.1109/TCT.1971.1083337.

[19] S. Kvatinsky *et al.*, « MAGIC—Memristor-Aided Logic », *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, nº 11, p. 895-899, nov. 2014, doi: 10.1109/TCSII.2014.2357292.

[20] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, et R. S. Williams, « 'Memristive' switches enable 'stateful' logic operations via material implication », *Nature*, vol. 464, nº 7290, p. 873-876, avr. 2010, doi: 10.1038/nature08940.

[21] S. Kvatinsky, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, et U. C. Weiser, « Memristor-Based Material Implication (IMPLY) Logic: Design Principles and Methodologies », *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, nº 10, p. 2054-2066, oct. 2014, doi: 10.1109/TVLSI.2013.2282132.

[22] T. Zanotti, F. M. Puglisi, et P. Pavan, « Smart Logic-in-Memory Architecture for Low-Power Non-Von Neumann Computing », *IEEE Journal of the Electron Devices Society*, vol. 8, p. 757-764, 2020, doi: 10.1109/JEDS.2020.2987402.

[23] S. M. Laube et N. TaheriNejad, « Device Variability Analysis for Memristive Material Implication ». Consulté le: 24 octobre 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/2101.07231

[24] L. Chua, « Five non-volatile memristor enigmas solved », *Applied Physics A*, vol. 124, nº 8, p. 563, juill. 2018, doi: 10.1007/s00339-018-1971-0.

[25] W.-H. Chen *et al.*, « A 16 Mb dual-mode ReRAM macro with sub-14 ns computing-in-memory and memory functions enabled by self-write termination scheme », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 28. doi: 10.1109/IEDM.2017.8268468.

[26] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, et Y. Xie, « Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories », in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, juin 2016, p. 1-6. doi: 10.1145/2897937.2898064.

[27] E. Esmanhotto *et al.*, « Experimental demonstration of Single-Level and Multi-Level-Cell RRAM-based In-Memory Computing with up to 16 parallel operations », in *2 022 IEEE International Reliability Physics Symposium (IRPS)*, mars 2022, p. P8–1–P8–4. doi: 10.1109/IRPS48227.2022.9764474.

[28] J. Chen, J. Li, Y. Li, et X. Miao, « Multiply accumulate operations in memristor crossbar arrays for analog computing », *Journal of Semiconductors*, vol. 42, nº 1, p. 13104, janv. 2021, doi: 10.1088/1674-4926/42/1/013104.

[29] J. M. Correll *et al.*, « An 8-bit 20.7 TOPS/W Multi-Level Cell ReRAM-based Compute Engine », in *2 022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, juin 2022, p. 264-265. doi: 10.1109/VLSITechnologyandCir46769.2022.9830490.

[30] W. Wan *et al.*, « 33.1 A 74 TMACS/W CMOS-RRAM Neurosynaptic Core with Dynamically Reconfigurable Dataflow and In-situ Transposable Weights for Probabilistic Graphical Models », in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, févr. 2020, p. 498-500. doi: 10.1109/ISSCC19947.2020.9062979.

[31] H. Jiang, W. Li, S. Huang, et S. Yu, « A 40 nm Analog-Input ADC-Free Compute-in-Memory RRAM Macro with Pulse-Width Modulation between Sub-arrays », in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, juin 2022, p. 266-267. doi: 10.1109/VLSITechnologyandCir46769.2022.9830211.

[32] A. Glukhov *et al.*, « Statistical model of program/verify algorithms in resistive-switching memories for in-memory neural network accelerators », in *2022 IEEE International Reliability Physics Symposium (IRPS)*, mars 2022, p. 3. doi: 10.1109/IRPS48227.2022.9764497.

[33] V. Milo *et al.*, « Multilevel HfO2-based RRAM devices for low-power neuromorphic networks », *APL Materials*, vol. 7, nº 8, p. 81120, août 2019, doi: 10.1063/1.5108650.

[34] C.-X. Xue *et al.*, « A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices », *Nature Electronics*, vol. 4, nº 1, p. 81-90, janv. 2021, doi: 10.1038/s41928-020-00505-5.

[35] M. Hu *et al.*, « Memristor-Based Analog Computation and Neural Network Classification with a Dot Product Engine », *Advanced Materials*, vol. 30, nº 9, p. 1705914, 2018, doi: 10.1002/adma.201705914.

[36] V. Milo *et al.*, « Accurate Program/Verify Schemes of Resistive Switching Memory (RRAM) for In-Memory Neural Network Circuits », *IEEE Transactions on Electron Devices*, vol. 68, nº 8, p. 3832-3837, août 2021, doi: 10.1109/TED.2021.3089995.

[37] S. Yu, « Neuro-inspired computing with emerging nonvolatile memorys », *Proceedings of the IEEE*, vol. 106, nº 2, p. 260-285, févr. 2018, doi: 10.1109/JPROC.2018.2790840.

[38] Y. Zhang *et al.*, « Optimized Programming Scheme Enabling Symmetric Conductance Modulation in HfO_2 Resistive Random-Access Memory (RRAM) for Neuromorphic Systems », *IEEE Electron Device Letters*, vol. 43, nº 8, p. 1203-1206, août 2022, doi: 10.1109/LED.2022.3182494.

[39] M. Bocquet *et al.*, « In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks », in *2018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 20. doi: 10.1109/IEDM.2018.8614639.

[40] C. Finn, P. Abbeel, et S. Levine, « Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks », *CoRR*, 2017, [En ligne]. Disponible sur: http://arxiv.org/abs/1703.03400

[41] M. Pallo *et al.*, « On Chip Customized Learning on Resistive Memory Technology for Secure Edge AI », in *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2025, p. 1-3. doi: 10.23919/VLSITechnologyandCir65189.2025.11074789.

[42] M. A. Zidan *et al.*, « A general memristor-based partial differential equation solver », *Nature Electronics*, vol. 1, nº 7, p. 411-420, juill. 2018, doi: 10.1038/s41928-018-0100-6.

[43] C.-X. Xue *et al.*, « 24.1 A 1 Mb Multibit ReRAM Computing-In-Memory Macro with 14,6 ns Parallel MAC Computing Time for CNN Based AI Edge Processors », in *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, févr. 2019, p. 388-390. doi: 10.1109/ISSCC.2019.8662395.

[44] C. Li *et al.*, « Analogue signal and image processing with large memristor crossbars », *Nature Electronics*, vol. 1, nº 1, p. 52-59, janv. 2018, doi: 10.1038/s41928-017-0002-z.

[45] K.-J. Zhou *et al.*, « Abnormal High Resistive State Current Mechanism Transformation in Ti/HfO2/TiN Resistive Random Access Memory », *IEEE Electron Device Letters*, vol. 41, nº 2, p. 224-227, févr. 2020, doi: 10.1109/LED.2019.2961408.

[46] R. Waser, R. Dittmann, G. Staikov, et K. Szot, « Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges », *Advanced Materials*, vol. 21, nº 25–26, p. 2632-2663, 2009, doi: 10.1002/adma.200900375.

[47] H.-S. P. Wong *et al.*, « Metal–Oxide RRAM », *Proceedings of the IEEE*, vol. 100, nº 6, p. 1951-1970, juin 2012, doi: 10.1109/JPROC.2012.2190369.

[48] F. M. Puglisi, N. Zagni, L. Larcher, et P. Pavan, « Random Telegraph Noise in Resistive Random Access Memories: Compact Modeling and Advanced Circuit Design », *IEEE Transactions on Electron Devices*, vol. 65, nº 7, p. 2964-2972, juill. 2018, doi: 10.1109/TED.2018.2833208.

[49] D. Ielmini, F. Nardi, et S. Balatti, « Evidence for Voltage-Driven Set/Reset Processes in Bipolar Switching RRAM », *IEEE Transactions on Electron Devices*, vol. 59, nº 8, p. 2049-2056, août 2012, doi: 10.1109/TED.2012.2199497.

[50] F. Nardi, S. Larentis, S. Balatti, D. C. Gilmer, et D. Ielmini, « Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part I: Experimental Study », *IEEE Transactions on Electron Devices*, vol. 59, nº 9, p. 2461-2467, sept. 2012, doi: 10.1109/TED.2012.2202319.

[51] Z.-Q. Wang *et al.*, « Cycling-induced degradation of metal-oxide resistive switching memory (RRAM) », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 7. doi: 10.1109/IEDM.2015.7409649.

[52] D. Ielmini, « Resistive switching memories based on metal oxides: mechanisms, reliability and scaling », *Semiconductor Science and Technology*, vol. 31, nº 6, p. 63002, juin 2016, doi: 10.1088/0268-1242/31/6/063002.

[53] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, et D. Ielmini, « Statistical Fluctuations in HfOx Resistive-Switching Memory: Part I - Set/Reset Variability », *IEEE Transactions on Electron Devices*, vol. 61, nº 8, p. 2912-2919, août 2014, doi: 10.1109/TED.2014.2330200.

[54] A. Grossi *et al.*, « Fundamental variability limits of filament-based RRAM », in *2 016 IEEE International Electron Devices Meeting (IEDM)*, déc. 2016, p. 4. doi: 10.1109/IEDM.2016.7838348.

[55] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, et D. Ielmini, « Statistical Fluctuations in HfOx Resistive-Switching Memory: Part II—Random Telegraph Noise », *IEEE Transactions on Electron Devices*, vol. 61, nº 8, p. 2920-2927, août 2014, doi: 10.1109/TED.2014.2330202.

[56] S. Ambrogio, S. Balatti, V. McCaffrey, D. C. Wang, et D. Ielmini, « Noise-Induced Resistance Broadening in Resistive Switching Memory—Part I: Intrinsic Cell Behavior », *IEEE Transactions on Electron Devices*, vol. 62, nº 11, p. 3805-3811, nov. 2015, doi: 10.1109/TED.2015.2475598.

[57] S. Ambrogio, S. Balatti, V. McCaffrey, D. C. Wang, et D. Ielmini, « Noise-Induced Resistance Broadening in Resistive Switching Memory—Part II: Array Statistics », *IEEE Transactions on Electron Devices*, vol. 62, nº 11, p. 3812-3819, nov. 2015, doi: 10.1109/TED.2015.2477135.

[58] L. Reganaz *et al.*, « Elucidating Postprogramming Relaxation in Multilevel Cell-Resistive Random Access Memory by Means of Experimental and Kinetic Monte Carlo Simulation Data », *physica status solidi (a)*, nº n/a, p. 2100753, doi: 10.1002/pssa.202100753.

[59] C. Y. Chen *et al.*, « Programming-conditions solutions towards suppression of retention tails of scaled oxide-based RRAM », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 10. doi: 10.1109/IEDM.2015.7409671.

[60] J. Sandrini *et al.*, « OxRAM for embedded solutions on advanced node: scaling perspectives considering statistical reliability and design constraints », in *2 019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 30. doi: 10.1109/IEDM19573.2019.8993484.

[61] J. Meza, Q. Wu, S. Kumar, et O. Mutlu, « Revisiting Memory Errors in Large-Scale Production Data Centers: Analysis and Modeling of New Trends from the Field », in *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, juin 2015, p. 415-426. doi: 10.1109/DSN.2015.57.

[62] H. Tang et J. Park, « Unequal-Error-Protection Error Correction Codes for the Embedded Memories in Digital Signal Processors », *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, nº 6, p. 2397-2401, juin 2016, doi: 10.1109/TVLSI.2015.2497368.

[63] C. Li, R. M. Roth, C. Graves, X. Sheng, et J. P. Strachan, « Analog error correcting codes for defect tolerant matrix multiplication in crossbars », in *2 020 IEEE International Electron Devices Meeting (IEDM)*, déc. 2020, p. 36. doi: 10.1109/IEDM13553.2020.9371978.

[64] R. M. Roth, « Analog Error-Correcting Codes », *IEEE Transactions on Information Theory*, vol. 66, nº 7, p. 4075-4088, juill. 2020, doi: 10.1109/TIT.2020.2977918.

[65] M. Jiang, R. Mao, J. P. Strachan, et C. Li, « Defect tolerant in-memory analog computing with CMOS-integrated nanoscale crossbars: Invited », in *2021 28th IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, nov. 2021, p. 1-4. doi: 10.1109/ICECS53924.2021.9665649.

[66] B. Chen *et al.*, « A Novel Operation Scheme for Oxide-Based Resistive-Switching Memory Devices to Achieve Controlled Switching Behaviors », *IEEE Electron Device Letters*, vol. 32, nº 3, p. 282-284, mars 2011, doi: 10.1109/LED.2010.2101577.

[67] W. Lian *et al.*, « Improved Resistive Switching Uniformity in $ \hboxCu/HfO_2/\hboxPt$ Devices by Using Current Sweeping Mode », *IEEE Electron Device Letters*, vol. 32, nº 8, p. 1053-1055, août 2011, doi: 10.1109/LED.2011.2157990.

[68] Y. Luo, X. Han, Z. Ye, H. Barnaby, J.-S. Seo, et S. Yu, « Array-Level Programming of 3-Bit per Cell Resistive Memory and Its Application for Deep Neural Network Inference », *IEEE Transactions on Electron Devices*, vol. 67, nº 11, p. 4621-4625, nov. 2020, doi: 10.1109/TED.2020.3015940.

[69] E. Esmanhotto *et al.*, « High-Density 3D Monolithically Integrated Multiple 1T1R Multi-Level-Cell for Neural Networks », in *2 020 IEEE International Electron Devices Meeting (IEDM)*, déc. 2020, p. 36. doi: 10.1109/IEDM13553.2020.9372019.

[70] Y. Feng *et al.*, « Improvement of State Stability in Multi-Level Resistive Random-Access Memory (RRAM) Array for Neuromorphic Computing », *IEEE Electron Device Letters*, vol. 42, nº 8, p. 1168-1171, août 2021, doi: 10.1109/LED.2021.3091995.

[71] B. Q. Le *et al.*, « RADAR: A Fast and Energy-Efficient Programming Technique for Multiple Bits-Per-Cell RRAM Arrays », *IEEE Transactions on Electron Devices*, vol. 68, nº 9, p. 4397-4403, sept. 2021, doi: 10.1109/TED.2021.3097975.

[72] A. Fantini *et al.*, « Intrinsic program instability in HfO2 RRAM and consequences on program algorithms », in *2 015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 7. doi: 10.1109/IEDM.2015.7409648.

[73] C.-Y. Wu *et al.*, « Emerging Memory RRAM Embedded in 12FFC FinFET Technology for industrial Applications », in *2 023 International Electron Devices Meeting (IEDM)*, déc. 2023, p. 1-4. doi: 10.1109/IEDM45741.2023.10413775.

[74] C. Peters, F. Adler, K. Hofmann, et J. Otterstedt, « Reliability of 28 nm embedded RRAM for consumer and industrial products », in *2 022 IEEE International Memory Workshop (IMW)*, mai 2022, p. 1-3. doi: 10.1109/IMW52921.2022.9779300.

[75] S. Amer, M. S. Hasan, et G. S. Rose, « Analysis and Modeling of Electroforming in Transition Metal Oxide-Based Memristors and Its Impact on Crossbar Array Density », *IEEE Electron Device Letters*, vol. 39, nº 1, p. 19-22, janv. 2018, doi: 10.1109/LED.2017.2778639.

[76] M. Barlas *et al.*, « Improvement of HfO2 based RRAM array performances by local Si implantation », in *2 017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 14. doi: 10.1109/IEDM.2017.8268392.

[77] L. Grenouillet *et al.*, « 16kbit 1T1R OxRAM arrays embedded in 28 nm FDSOI technology demonstrating low BER, high endurance, and compatibility with core logic transistors », in *2 021 IEEE International Memory Workshop (IMW)*, mai 2021, p. 1-4. doi: 10.1109/IMW51353.2021.9439607.

[78] S. Zhang *et al.*, « Programming resistive switching memory by a charged capacitor », *Applied Physics A*, vol. 102, nº 4, p. 1003-1007, mars 2011, doi: 10.1007/s00339-011-6320-5.

[79] P. Trotti, G. Pillonet, G. Molas, S. Oukassi, et E. Nowak, « Experimental Set-Up For Novel Energy Efficient Charge-based Resistive RAM (RRAM) Switching », in *2 020 IEEE 33rd International Conference on Microelectronic Test Structures (ICMTS)*, mai 2020, p. 1-5. doi: 10.1109/ICMTS48187.2020.9107936.

[80] A. Padovani, J. Woo, H. Hwang, et L. Larcher, « Understanding and Optimization of Pulsed SET Operation in HfOx-Based RRAM Devices for Neuromorphic Computing Applications », *IEEE Electron Device Letters*, vol. 39, nº 5, p. 672-675, mai 2018, doi: 10.1109/LED.2018.2821707.

[81] J. Reuben, M. Biglari, et D. Fey, « Incorporating Variability of Resistive RAM in Circuit Simulations Using the Stanford–PKU Model », *IEEE Transactions on Nanotechnology*, vol. 19, p. 508-518, 2020, doi: 10.1109/TNANO.2020.3004666.

[82] P. R. Shrestha *et al.*, « Analysis and Control of RRAM Overshoot Current », *IEEE Transactions on Electron Devices*, vol. 65, nº 1, p. 108-114, janv. 2018, doi: 10.1109/TED.2017.2776860.

[83] P. Trotti, « Dispositifs et circuits à convenance énergétique pour les mémoires résistives », 2022. [En ligne]. Disponible sur: http://www.theses.fr/2022GRALY073/document

[84] C. Cagli *et al.*, « Study of the Energy Consumption Optimization on RRAM Memory Array for SCM Applications », in *2 017 IEEE International Memory Workshop (IMW)*, 2017, p. 1-4. doi: 10.1109/IMW.2017.7939106.

[85] G. Piccolboni *et al.*, « Investigation of Cycle-to-Cycle Variability in HfO2-Based OxRAM », *IEEE Electron Device Letters*, vol. 37, nº 6, p. 721-723, juin 2016, doi: 10.1109/LED.2016.2553370.

[86] X. Xu *et al.*, « 40× Retention Improvement by Eliminating Resistance Relaxation with High Temperature Forming in 28 nm RRAM Chip », in *2 018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 20. doi: 10.1109/IEDM.2018.8614593.

[87] E. Perez, M. K. Mahadevaiah, E. P.-B. Quesada, et C. Wenger, « Variability and Energy Consumption Tradeoffs in Multilevel Programming of RRAM Arrays », *IEEE Transactions on Electron Devices*, vol. 68, nº 6, p. 2693-2698, juin 2021, doi: 10.1109/TED.2021.3072868.

[88] « Cypress eCT Flash: The Most Scalable High-Performance, High-Reliability Embedded Flash Solution for Next-Generation MCUs », in *Infineon Product Brief "eCT Flash"*, 2021.

[89] J. Cooke, « NAND Flash 101: An Introduction to NAND Flash and How to Design It in to Your Next Product », in *Micron Technical Note 29 19 (TN-29-19)*, 2016.

[90] A. Lu, X. Peng, W. Li, H. Jiang, et S. Yu, « NeuroSim Simulator for Compute-in-Memory Hardware Accelerator: Validation and Benchmark », *Frontiers in Artificial Intelligence*, vol. 4, p. , 2021, doi: 10.3389/frai.2021.659060.

[91] T. Liu, S. Zhu, M. H. White, A. Salemi, D. Sheridan, et A. K. Agarwal, « Time-Dependent Dielectric Breakdown of Commercial 1.2 kV 4H-SiC Power MOSFETs », *IEEE Journal of the Electron Devices Society*, vol. 9, p. 633-639, 2021, doi: 10.1109/JEDS.2021.3091898.

[92] H.-X. Zheng *et al.*, « Reducing Forming Voltage by Applying Bipolar Incremental Step Pulse Programming in a 1T1R Structure Resistance Random Access Memory », *IEEE Electron Device Letters*, vol. 39, nº 6, p. 815-818, juin 2018, doi: 10.1109/LED.2018.2831708.

[93] D. A. Robayo *et al.*, « Integration of OTS based back-end selector with HfO2 OxRAM for crossbar arrays », in *2019 IEEE 11th International Memory Workshop (IMW)*, mai 2019, p. 1-4. doi: 10.1109/IMW.2019.8739746.

[94] J. Zhang *et al.*, « A 3D RRAM using stackable 1TXR memory cell for high density application », in *2009 International Conference on Communications, Circuits and Systems*, juill. 2009, p. 917-920. doi: 10.1109/ICCCAS.2009.5250369.

[95] J. Yang *et al.*, « A 28 nm 1,5 Mb Embedded 1T2R RRAM with 14.8 Mb/mm2 using Sneaking Current Suppression and Compensation Techniques », in *2020 IEEE Symposium on VLSI Circuits*, juin 2020, p. 1-2. doi: 10.1109/VLSICircuits18222.2020.9163035.

[96] C.-W. S. Yeh et S. S. Wong, « Compact One-Transistor-N-RRAM Array Architecture for Advanced CMOS Technology », *IEEE Journal of Solid-State Circuits*, vol. 50, nº 5, p. 1299-1309, mai 2015, doi: 10.1109/JSSC.2015.2402217.

[97] L. Zhao, L. Jiang, Y. Zhang, N. Xiao, et J. Yang, « Constructing fast and energy efficient 1TnR based ReRAM crossbar memory », in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, mars 2017, p. 58-64. doi: 10.1109/ISQED.2017.7918293.

[98] E. R. Hsieh *et al.*, « Four-Bits-Per-Memory One-Transistor-and-Eight-Resistive-Random-Access-Memory (1T8R) Array », *IEEE Electron Device Letters*, vol. 42, nº 3, p. 335-338, mars 2021, doi: 10.1109/LED.2021.3055017.

[99] E. Hsieh *et al.*, « High-Density Multiple Bits-per-Cell 1T4R RRAM Array with Gradual SET/RESET and its Effectiveness for Deep Learning », in *2019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 35. doi: 10.1109/IEDM19573.2019.8993514.

Liste de publications

[100] T. Bauvent *et al.*, « RRAM Device Performances Under Capacitive-Enhanced Current Programming Scheme », *IEEE Electron Device Letters*, vol. 44, nº 7, p. 1112-1115, juill. 2023, doi: 10.1109/LED.2023.3274219.

[101] T. Bauvent, G. Pillonnet, et G. Molas, « Optimizing RRAM Performance: A Comparative Analysis of Forming Strategies », in *2024 IEEE International Memory Workshop (IMW)*, mai 2024, p. 1-4. doi: 10.1109/IMW59701.2024.10536941.

[102] T. Bauvent, G. Pillonnet, et G. Molas, « How Significant Is SET Programming Strategy in Enhancing RRAM Technology? », in *2024 IEEE European Solid-State Electronics Research Conference (ESSERC)*, sept. 2024, p. 653-656. doi: 10.1109/ESSERC62670.2024.10719434.

Brevet en attente de publication : T. Bauvent, G. Pillonnet, G. Molas, « Dispositif mémoire à matrice de cellules mémoires », FR2412063 – A1.

Annexe A : Instructions disponibles sur la plateforme de tests du circuit *MEMORIIC*

Le Tableau A.1 liste, de manière exhaustive, les instructions disponibles dans un programme à envoyer au micro-contrôleur de la plateforme de tests décrite dans le Chapitre IV du circuit *MEMORIIC*. On rappelle que chaque instruction est encodée sur 32 bits (petit-boutisme).

Tableau A.1 – Tableau exhaustif des instructions disponibles sur le micro-contrôleur de la plateforme de tests du circuit *MEMORIIC*.

Nom	Encodage[α]				Description
	octet 0	octet 1	octet 2	octet 3	
Nop	0x00	–	–	–	Aucun effet
Num_Set_Bus	0x01	–	Valeur 14 bits		Définit la valeur du bus de 14 bits à la valeur indiquée
Num_Set_Bits	0x02	Port	Numéro de pin		Met à l'état actif la sortie numérique correspondant au port et au numéro de pin donnés
Num_Reset_Bits	0x03	Port	Numéro de pin		Met à l'état inactif la sortie numérique correspondant au port et au numéro de pin donnés
Dac_Set_Output	0x04	Id DAC	Valeur 12 bits		Définit la valeur numérique de la tension de sortie du DAC indiqué
Adc_Read	0x05	Id ADC	–	–	Commence et attend la fin d'une conversion d'ADC et stocke la valeur obtenue dans le tampon des résultats
Adc_Configure	0x06	Id ADC	T_{ech}	N_{ech}	Modifie la configuration de l'ADC
Adc_Placeholder_Result	0x0F	–	–	–	Stocke une valeur spéciale dans le tampon des résultats
Transfert_Results	0x10	Mode	–	–	Démarre le transfert des résultats dans le tampon des résultats vers l'ordinateur
Wait_Ticks	0x20	Nombre de coups			Attend pendant le nombre indiqué de coups d'horloge (du micro-processeur du micro-contrôleur)

Register_Set	0x40	Reg	Valeur		Définit la valeur du registre indiqué
Register_Add	0x41	Reg	Valeur		Ajoute à la valeur du registre indiqué la valeur indiquée
Register_Sub	0x42	Reg	Valeur		Soustrait à la valeur du registre indiqué la valeur indiquée
Set_Mask_If_Last_Read_Above_Threshold	0x4E	Reg	Seuil		Masque l'adresse ciblée par le registre indiqué si la dernière valeur ajoutée au tampon des résultats est supérieure au seuil indiqué
Set_Mask	0x4F	Reg	–	Bool	Masque ou démasque l'adresse ciblée par le registre indiqué
Jump	0x50	–	Valeur		Déplace le compteur d'instructions à la valeur indiquée
Jump_If_Zero	0x51	Reg	Valeur		Déplace le compteur d'instructions à la valeur indiquée si la valeur dans le registre indiqué est nulle
Jump_If_Not_Zero	0x52	Reg	Valeur		Déplace le compteur d'instructions à la valeur indiquée si la valeur dans le registre indiqué n'est pas nulle
Jump_If_Masked	0x53	Reg	Valeur		Déplace le compteur d'instructions à la valeur indiquée si l'adresse ciblée par le registre indiqué est masquée
Jump_If_Not_Masked	0x54	Reg	Valeur		Déplace le compteur d'instructions à la valeur indiquée si l'adresse ciblé par le registre indiqué n'est pas masquée
Stop_Point	0x60	–	–	–	Définit un point d'arrêt possible pour l'exécution
Stop_Execution	0x61	–	–	–	Signal de demande d'arrêt de l'exécution au prochain point d'arrêt. Peut être envoyé pendant l'exécution
Num_Set_Bus_R	0x81	–	–	Reg	Définit la valeur du bus de 14 bits à la valeur contenue dans le registre indiqué
Register_Set_R	0xC0	Reg (Dest)	Reg (Src)	–	Définit la valeur du registre (Dest) à la valeur du registre (Src)
Register_Add_R	0xC1	Reg (Dest)	Reg (Op1)	Reg (Op2)	Définit la valeur du registre (Dest): (Dest) <- (Op1) + (Op2)
Register_Sub_R	0xC2	Reg (Dest)	Reg (Op1)	Reg (Op2)	Définit la valeur du registre (Dest): (Dest) <- (Op1) - (Op2)
Register_Load_R	0xC3	Reg (Dest)	Reg (iSrc)	–	Copie la valeur du registre indexé par la valeur du registre (iSrc) dans le registre (Dest)

[α] L'encodage est petit-boutisme, l'octet 0 est le premier octet dans l'ordre mémoire, donc celui de point faible. Lorsque plusieurs octets sont combinés pour former une valeur 16 bits, son encodage est également petit-boutisme.

Annexe B : Détails du circuit *CAPA-WR*

B.1 Pads

Le Tableau B.1 décrit les 25 pads du circuit *CAPA-WR*.

Tableau B.1 – Descriptif des plots d'accès du circuit *CAPA-WR*

Index	Nom	Type	Description
1	vdd	Alimentation	Alimentation 4,6 V nominal.
2	gnd	Alimentation	Masse électrique.
3	cell_wl	Analogique	Tension appliquée à la grille du sélecteur sélectionné par ADDRESS<0–13> pendant l'écriture (lorsque enable_write est actif).
4	cell_low	Analogique	Tension appliquée à la borne de plus faible potentiel de la cellule 1T1R, c'est-à-dire à la ligne de bit, respectivement la ligne de source, sélectionnée lorsque IS_RESET est actif, respectivement est inactif. Généralement connecté à la masse gnd
5	unselected_cell	Analogique	Tension appliquée aux lignes de bit et lignes de sources non sélectionnées.
6	cell_high	Analogique	Tension appliquée à la borne de plus fort potentiel de la cellule 1T1R, c'est-à-dire à la ligne de bit, respectivement la ligne de source, sélectionnée lorsque IS_RESET est inactif, respectivement est actif.
7	enable_write	Entrée numérique	LOW : Écriture/lecture inactive. Aucun sélecteur n'est passant. HIGH : Écriture/lecture active. La grille du sélecteur adressé par ADDRESS<0–13> est connectée à cell_wl.
8–21	memory_in<0–13>	Entrée numérique	Le bus numérique permettant de définir les valeurs des registres intégrés.
22–23	memory_address<0–1>	Entrée numérique	L'adresse du registre où copier memory_in<0–13>

Index	Nom	Type	Description
24	lock_memory	Entrée numérique	LOW : memory_in<0–13> est recopié dans le registre adressé par memory_address<0–1>.
			HIGH : Aucun des registres n'est modifié.
25	auto_reload_clk	Horloge	Horloge de référence pour le mode capacitif rechargé. Utile uniquement lorsque AUTO_RELOAD est actif.

B.2 Mémoire de configuration

La mémoire de configuration est accessible au travers du bus d'adresse de 2 bits memory_address<0–1>, du bus de données de 14 bits memory_in<0–13> également utilisé pour adresser la matrice d'éléments mémoires OxRAM et du bit lock_memory contrôlant l'écriture ou non du registre. Les registres sont composés chacun de 14 bits et sont au nombre de 4. Le Tableau B.2 ainsi que le Tableau B.3 détaillent le contenu de ces registres.

La Figure B.1 illustre une séquence d'écriture d'un registre. lock_memory est d'abord conservé HIGH. La valeur du bus memory_in<0–13> ainsi que l'adresse du registre memory_address<0–1> sont fixées à leur valeur souhaitée. Enfin, lock_memory est mis LOW, puis remonté au 1 logique. La valeur de memory_in<0–13> et memory_address<0–1> peut alors être changée sans risque.

Après l'alimentation du circuit, l'état initial de la mémoire n'est pas certain, il est donc préférable d'écrire les 4 registres avant toute autre opération. La valeur des registres ne peut pas être lue et ses 14 bits doivent être écrits simultanément.

Tableau B.2 – Descriptif des registres du circuit *CAPA-WR*

		memory_address<0–1> au format <a1 a0>			
		00	01	10	11
memory_in<0–13>	b0	ADDRESS<0>	IS_RESET	AUTO_RELOAD	MOM<0>
	b1	⋮	MODE	AUTO_END	⋮
	b2	⋮	CAPA_EXT	TH_HIGH<0>	⋮
	b3	⋮	DOUBLE	⋮	⋮
	b4	⋮	BANK1	⋮	⋮
	b5	⋮	PRE_ADDR	⋮	⋮
	b6	⋮	MEM_CAPA	TH_HIGH<4>	MOM<6>
	b7	⋮	MESURE_CAPA_EXT	TH_LOW<0>	MOS<0>
	b8	⋮	⋮	⋮	⋮
	b9	⋮	⋮	⋮	⋮
	b10	⋮	⋮	⋮	⋮
	b11	⋮	⋮	TH_LOW<4>	⋮
	b12	⋮	⋮	TH_HIGH_USE_PMOS	⋮
	b13	ADDRESS<13>	⋮	TH_LOW_USE_PMOS	MOS<6>

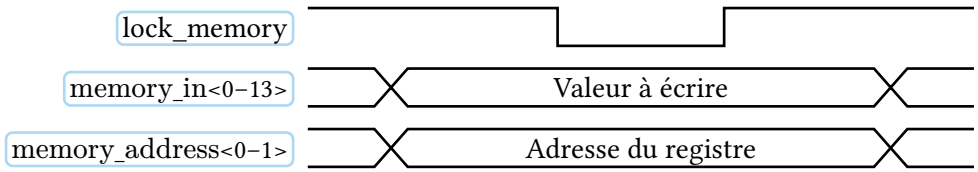


Figure B.1 – Chronogramme de l'écriture dans la valeur « Valeur à écrire » dans le registre à l'adresse « Adresse du registre ».

Tableau B.3 – Descriptif des entrées des registres du circuit *CAPA-WR*

Nom	État	Description
ADDRESS<0–13>	–	L'adresse de l'élément mémoire OxRAM sélectionné
IS_RESET	0	SET : La ligne de bit sélectionnée est connectée à cell_high_int. La ligne de source sélectionnée est connectée à cell_low.
	1	RESET : La ligne de bit sélectionnée est connectée à cell_low. La ligne de source sélectionnée est connectée à cell_high_int.
MODE	0	Mode tension : Quand enable_write est actif, connecte les signaux cell_high et cell_low à l'élément mémoire OxRAM sélectionné.
	1	Mode capacitif : Quand enable_write est actif, connecte la capacité sélectionnée à l'élément mémoire OxRAM sélectionné. Quand enable_write est inactif, connecte cell_high à la capacité sélectionnée pour la charger.
CAPA_EXT	0	Avec MODE au 1 logique : cell_high est isolé de cell_high_int lorsque enable_write est actif.
	1	Avec MODE au 1 logique : cell_high reste connecté à cell_high_int lorsque enable_write est actif. Permet l'observation du signal et l'ajout de capacité externe. (Incompatible avec AUTO_RELOAD)
DOUBLE	0	–
	1	Utilise les deux banques de capacité à la fois ce qui double la capacité effective. (Incompatible avec AUTO_RELOAD)
BANK1	0	Utilise la banque de capacités 0.
	1	Utilise la banque de capacités 1.
PRE_ADDR	0	Tant que enable_write est inactif, toutes les lignes de bit et de source de la matrice sont connectées à unselected_cell. Dès que enable_write est actif, la ligne de bit, respectivement de source, est connectée à cell_high_int, respectivement cell_low.
	1	La ligne de bit, respectivement de source, sélectionnée est toujours connectée à cell_high_int, respectivement cell_low.
MEM_CAPA	0	–
	1	Les lignes de bit et de source non sélectionnées sont connectées à cell_high_int. Cela permet d'ajouter la capacité provenant de la construction de la matrice d'éléments mémoires OxRAM.
MESURE_CAPA_EXT	0	–
	1	Déconnecte cell_high de cell_high_int indépendamment de enable_write et CAPA_EXT.

Nom	État	Description
AUTO_RELOAD	0	–
	1	Mode avec recharge : Pendant que enable_write est actif, si cell_high_int est compris entre TH_LOW<0–4> et TH_HIGH<0–4> pendant deux coups d'horloge consécutifs de auto_reload_clk, la banque de capacité utilisée est alternée. (Incompatible avec CAPA_EXT et DOUBLE)
AUTO_END	0	–
	1	Pendant que enable_write est actif, si cell_high_int est inférieur à TH_LOW<0–4>, alors enable_write est tiré vers gnd par une résistance de 1 kΩ.
TH_HIGH<0–4>	–	La valeur du seuil haut pour le mode avec recharge. Soit x la valeur numérique de TH_HIGH<0–4>, la valeur du seuil est 0,6 V + x ∗ 0, 125 V. (min : 0,6 V, max : 4, 475 V)
TH_LOW<0–4>	–	La valeur du seuil bas pour le mode avec recharge. Soit x la valeur numérique de TH_LOW<0–4>, la valeur du seuil est 0,6 V + x ∗ 0, 125 V. (min : 0,6 V, max : 4, 475 V)
TH_HIGH_USE_PMOS	0	La paire différentielle d'entrée du comparateur pour le seuil haut utilise des NMOS. (Conseillé pour TH_HIGH<0–4> ≥ 2 V)
	1	La paire différentielle d'entrée du comparateur pour le seuil haut utilise des PMOS. (Conseillé pour TH_HIGH<0–4> < 2 V)
TH_LOW_USE_PMOS	0	La paire différentielle d'entrée du comparateur pour le seuil bas utilise des NMOS. (Conseillé pour TH_LOW<0–4> ≥ 2 V)
	1	La paire différentielle d'entrée du comparateur pour le seuil bas utilise des PMOS. (Conseillé pour TH_LOW<0–4> < 2 V)
MOM<0–6>	–	La valeur de capacité MOM à utiliser. La valeur unitaire est 0,6 pF pour un maximum à 76,2 pF.
MOS<0–6>	–	La valeur de capacité MOS à utiliser. La valeur unitaire est 3,8 pF pour un maximum à 482,6 pF.

B.3 Matrice mémoire

Les contrôles relatifs à la matrice d'éléments mémoires OxRAM ainsi que sa connexion au reste du circuit sont :

1. IS_RESET, le choix de la direction d'écriture;
2. ADDRESS<0–13>, la sélection de l'adresse;
3. enable_write, l'activation de la matrice connectant l'élément mémoire OxRAM sélectionné au reste du circuit;
4. cell_wl, la tension de grille des transistors servant de sélecteur au sein de la matrice d'éléments mémoires OxRAM et éventuellement de limitateur de courant;
5. unselected_cell, la tension appliquée sur les lignes et colonnes non-sélectionnées de la matrice;
6. PRE_ADDR, le contrôle de la pré-charge des nœuds internes de la matrice pour éventuellement limiter le phénomène de partage de charges;
7. MEM_CAPA, définit si la capacité interne de la matrice doit être utilisée comme capacité d'écriture (en plus de la capacité d'écriture sélectionnée par MOM<0–6> et MOS<0–6>)

B.4 Mode écriture en tension

Pour effectuer une écriture en tension, il faut :

1. Sélectionner le mode tension via MODE = 0 et la direction d'écriture via IS_RESET ;
2. Effectuer les autres configurations relatives à la matrice d'éléments mémoires OxRAM
3. Imposer la tension souhaitée entre cell_low et cell_high ainsi que la tension cell_wl à l'aide de sources externes;
4. Fixer enable_write HIGH pendant la durée de l'écriture souhaitée, puis le remettre LOW.

La Figure B.2 illustre une écriture en tension.

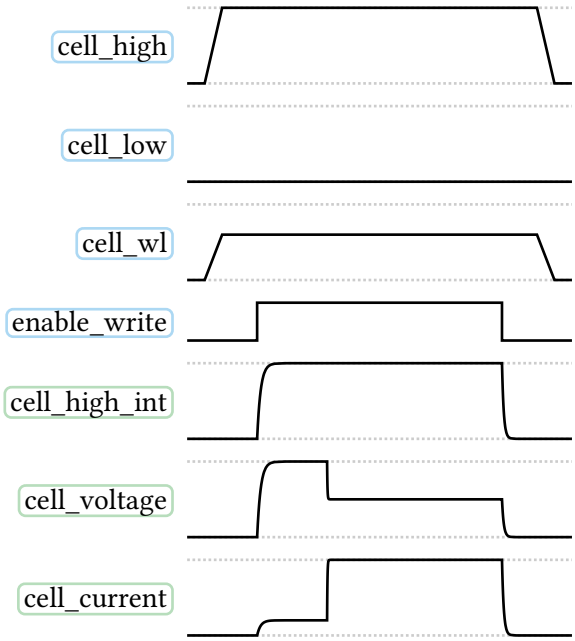


Figure B.2 – Chronogramme d'une écriture en tension.

B.5 Mode de lecture

Le mode de lecture est semblable au mode d'écriture en tension. La configuration est identique. C'est à l'utilisateur de veiller à imposer une tension entre cell_low et cell_high suffisamment faible pour ne pas provoquer d'écriture. Après avoir mis enable_write HIGH, le courant circulant entre cell_low et cell_high doit être mesuré depuis l'extérieur du circuit.

La Figure B.3 illustre une lecture.

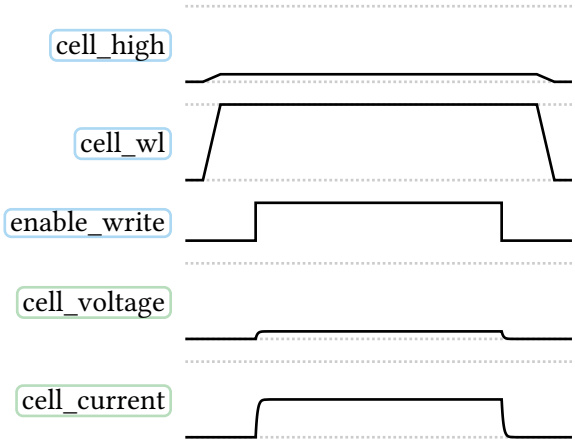


Figure B.3 – Chronogramme d'une lecture.

B.6 Méthode de mesure des capacités intégrées

L'architecture du circuit permet de mesurer les capacités intégrées depuis l'extérieur.

1. La première étape est de mesurer la capacité externe présente sur le chemin de mesure que le circuit ne voit pas en fonctionnement normal. MESURE_CAPA_EXT permet d'effectuer cette mesure en déconnectant cell_high des nœuds internes au circuit.
2. On peut ensuite mesurer la capacité pour différentes valeurs de MOM<0–6> et MOS<0–6> avec potentiellement MEM_CAPA
3. On peut en déduire la capacité vue par le circuit en fonctionnement en soustrayant les deux.

B.7 Mode d'écriture capacitive sans recharge

Pour effectuer une écriture capacitive sans recharge, il faut :

1. Sélectionner le mode tension via MODE = 1 et la direction d'écriture via IS_RESET;
2. Choisir la quantité de capacité MOM via MOM<0–6> (un pas de 0,6 pF jusqu'à 76,2 pF);
3. Choisir la quantité de capacité MOS via MOS<0–6> (un pas de 3,8 pF jusqu'à 482,6 pF);
4. Sélection de la ou des banques utilisées via BANK1 et DOUBLE
5. Choisir l'utilisation de la capacité interne de la matrice comme capacité d'écriture MEM_CAPA, en plus des capacités d'écriture sélectionnées par MOM<0–6> et MOS<0–6>;
6. Utiliser la pré-charge pour éviter la chute de tension due au partage de charges via PRE_ADDR (optionnel);
7. Éventuellement, si l'on veut ajouter une capacité externe, ou suivre l'évolution de la tension sur la capacité, CAPA_EXT permet de conserver la connexion entre cell_high et le signal interne lorsque enable_write est actif;
8. Imposer la tension de charge de la capacité sur cell_high ainsi que la tension cell_wl et mettre cell_low à GND;
9. Fixer enable_write HIGH pendant la durée de l'écriture souhaitée, puis le remettre LOW.

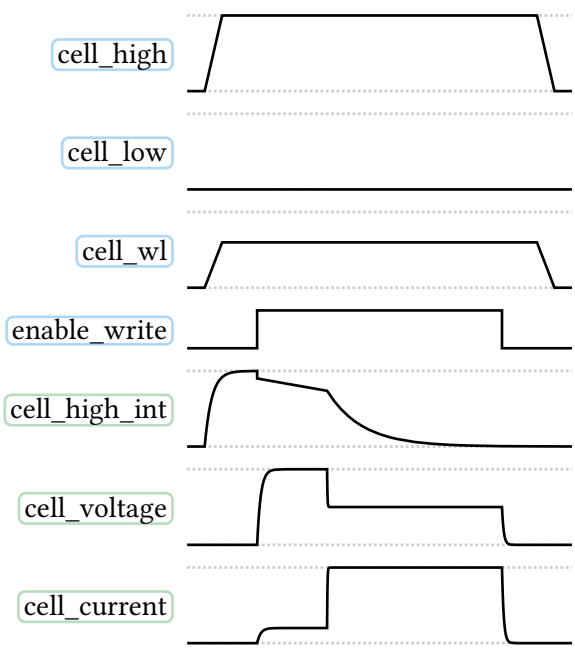


Figure B.4 – Chronogramme d'une écriture capacitive.

La Figure B.4 illustre une écriture capacitive sans recharge

B.8 Mode d'écriture capacitive avec recharge

Pour effectuer une écriture capacitive avec recharge, c'est la même méthode que pour l'écriture capacitive sans recharge avec en plus:

1. L'activation du mode avec recharge via AUTO_RELOAD;
2. L'application d'une horloge d'une fréquence choisie par l'utilisateur sur auto_reload_clk (la fréquence affecte la vitesse de réaction et les seuils de détection en terme de résistance);
3. Le paramétrage des seuils encadrant la zone de rechargement avec TH_HIGH<0–4> et TH_LOW<0–4> qui ont un pas de 125 mV, une valeur minimum de 0,6 V et maximum de 4,475 V;
4. La sélection du mode de fonctionnement des comparateurs avec TH_HIGH_USE_PMOS et TH_LOW_USE_PMOS qui devrait être actif pour tout seuil inférieur à 2 V et LOW sinon.

Les seuils TH_HIGH<0–4> et TH_LOW<0–4> ainsi que la fréquence de l'horloge auto_reload_clk doivent être réglés correctement pour que la résistance de l'élément mémoire OxRAM soit supérieure à $2R_0$ en HRS et soit inférieure à R_0 en LRS avec R_0 est définie par (B.1), en notant T la période de l'horloge auto_reload_clk et C la capacité d'écriture,

$$R_0 = \frac{T}{2C}\left(\frac{1.2}{0.125} + \frac{\text{TH_HIGH<0–4>} + \text{TH_LOW<0–4>}}{\text{TH_HIGH<0–4>} - \text{TH_LOW<0–4>}}\right) \tag{B.1}$$

Si l'on souhaite pouvoir observer la durée d'écriture AUTO_END permet de tirer à la masse enable_write lorsque la fin de l'écriture est détectée. Cela affectera aussi l'écriture puisque la capacité ne finira pas de se vider dans la mémoire après la fin de l'écriture.

Annexe C : Preuve de la condition de séparabilité de $2^n - 1$ états avec la méthode de lecture relative sur une cellule 1TnR

L'utilisation des éléments mémoires OxRAM en tant que mémoire numérique ne requiert pas la mesure de la conductance de chaque élément mémoire OxRAM. Être capable de distinguer l'état de chaque éléments mémoires OxRAM suffit à pouvoir lire l'information stockée. En imposant une tension V_i sur le terminal libre de chaque éléments mémoires OxRAM au sein d'une cellule 1TnR et un courant nul dans le transistor (en présentant une forte impédance sur SL) (voir Figure C.1), l'état de conductance relatif des éléments mémoires OxRAM, noté a_i, peut être déterminé. En effet, le potentiel du terminal commun à tous les éléments mémoires OxRAM de la cellule V_0, qui est également celui de la SL, est égal à la moyenne des potentiels pondérée par la conductance de chaque élément mémoire OxRAM comme exprimé par (C.1). La valeur d'un a_i peut être déterminée en mesurant V_0 en choisissant $V_i \neq 0$ et les autres $V_j = 0$.

$$V_0 = \frac{\sum_{i=1}^{n} \gamma_i V_i}{\sum_{i=1}^{n} \gamma_i} = \sum_{i=1}^{n} a_i V_i \text{ avec } a_i = \frac{\gamma_i}{\sum_{k=1}^{n} \gamma_k} \tag{C.1}$$

Peu importe le nombre d'observations effectuées, et le choix de V_i utilisé, seules les conductances relatives a_i pourront être déterminées, les valeurs absolues des conductances γ_i ne sont pas mesurables de cette manière[52].

Les a_i ainsi définies par (C.1) respectent $\sum_{i=1}^{n} a_i = 1$ et correspondent donc aux coordonnées barycentriques de V_0 dans le repère affine formé par les V_i. Pour pouvoir déterminer l'état de la mémoire par l'observation de V_0, il est nécessaire que les lieux de V_0 correspondants à chaque état soient disjoints.

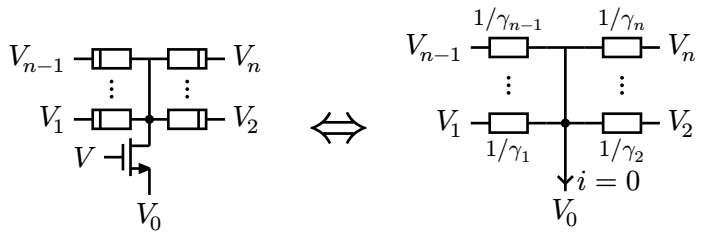


Figure C.1 – Cellule 1TnR sélectionnée (le sélecteur est passant) vue comme un réseau de résistance. Le courant i peut être observé à travers la SL.

[52]Les valeurs absolues des conductances γ_i ne peuvent pas être déterminées par simple observation de V_0. En effet, considérons deux situations, la première avec les conductances γ_i et la seconde avec les conductances $k\gamma_i$. Dans ces deux situations, les valeurs prises par V_0 en fonction des autres potentiels V_i seront identiques. Ainsi il est impossible de déterminer le facteur k, et seul le rapport relatif entre les γ_i nous est accessible.

Dans la suite, nous nous restreignons au cas où chaque élément mémoire OxRAM peut être dans l'un de deux états, LRS et HRS. De plus, nous considérons que la conductance d'un élément mémoire OxRAM en HRS appartient à l'intervalle $[0, h]$ et celle d'un élément en LRS à l'intervalle $[l, 1]$. Fixer ainsi le maximum de conductance à 1 ne perd pas en généralités puisque la valeur absolue des conductances est sans intérêt ici.

Il nous faut maintenant déterminer le lieu de V_0 dans chacun des états possibles, et plus particulièrement la condition sur h et l telle que ces lieux soient disjoints. Dans la suite, les différents états possibles de la cellule seront dénommés par une suite de H, pour HRS, et L, pour LRS, telle que chaque lettre indique l'état de l'élément mémoire OxRAM correspondant (en respectant l'ordre). Par exemple, l'état HLL correspond à la situation où $n = 3$ et $\gamma \in [0, h] \times [l, 1]^2$.

Pour commencer, notons que le lieu de V_0 est précisément l'enveloppe convexe des images des points extrêmes de l'ensemble des états de conductance correspondants à un état de la cellule. En effet, comme l'ensemble des conductances associé à un état est un produit cartésien d'intervalles de $\mathbb{R}$, il est convexe. Or la transformation reliant les conductances γ aux conductances relatives a conserve le caractère convexe des ensembles[53]. Donc les lieux de V_0 sont convexes. De plus, on peut montrer que tout antécédent d'un point extrême d'un lieu de V_0 est un point extrême de l'ensemble de départ[54], ce qui nous mène à notre précédente conclusion. A partir de cette propriété, on peut tracer les lieux de V_0. La Figure C.2 illustre le résultat ainsi obtenu pour $h = 0.2$, $l = 0.6$ pour $n = 2$, 3 et 4.

Nous explicitons la procédure d'obtention du lieu de V_0 à partir de la précédente propriété pour le cas $n = 2$. Les états possibles sont LL, LH, HL et HH. Dans l'état LL, les 4 points extrémaux de l'ensemble de départ sont $\binom{l}{l}$, $\binom{1}{l}$, $\binom{l}{1}$ et $\binom{1}{1}$, donnant respectivement $a_2 = \frac{1}{2}$, $\frac{l}{1+l}$, $\frac{1}{l+1}$ et $\frac{1}{2}$. La valeur de a_1 peut être déduite de a_2 grâce à la propritété $\sum_i a_i = 1$, seule la valeur de a_2 est donc rapportée. Le lieu de V_0 pour l'état LL est donc $\left[\frac{l}{1+l}, \frac{1}{l+1}\right]$. Pour l'état LH, les 4 points extrémaux de l'ensemble de départ sont $\binom{l}{0}$, $\binom{1}{0}$, $\binom{l}{h}$ et $\binom{1}{h}$, donnant respectivement $a_2 = 0, 0$, $\frac{h}{l+h}$ et $\frac{h}{1+h}$, ce qui donne $\left[0, \frac{h}{l+h}\right]$ comme lieu de V_0 pour LH. De même, on obtient $\left[\frac{l}{h+l}, 1\right]$ pour HL et $[0, 1]$ pour HH.

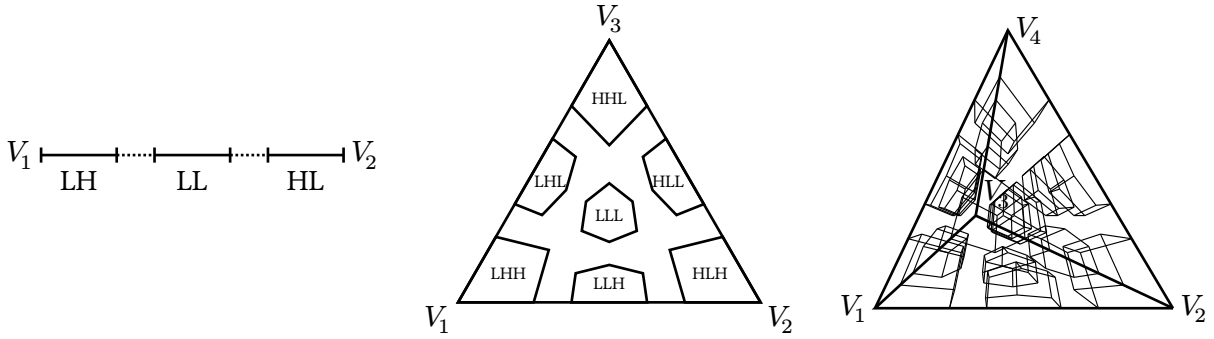


Figure C.2 – Représentation des lieux de V_0 pour $h = 0.2$ et $l = 0.6$ pour, de gauche à droite, $n = 2$, 3 et 4. Le lieu correspondant à l'état composé uniquement de H n'est pas représenté car il recouvrirait l'intégralité du simplexe.

[53] Soient $E \subset \mathbb{R}^n$ convexe, $F \subset \mathbb{R}^{n-1}$ tel que $\varphi : E \longmapsto F$ avec $\varphi(x) = \frac{1}{\sum_i x_i} x$ soit surjective. Soient $y^0, y^1 \in F$ et $t \in [0, 1]$, on note $y^t = ty^0 + (1-t)y^1$, x^0 et x^1 des antécédents respectifs de y^0 et y^1 par φ, $s^0 = \sum_i x_i^0$ et $s^1 = \sum_i x_i^1$. On pose $u = \frac{s^1 t}{s^0 + t(s^1 - s^0)}$ et remarquons que $u \in [0, 1]$. Comme E est convexe, $x^t = ux^0 + (1-u)x^1 \in E$. Donc $\varphi(x^t) = \frac{ux^0 + (1-u)x^1}{\sum_i ux^0 + (1-u)x^1} = \frac{ux^0 + (1-u)x^1}{us^0 + (1-u)s^1} = \frac{ts^1 x^0 + (s^0 + t(s^1 - s^0) - ts^1)x^1}{ts^0 s^1 + (s^0 + t(s^1 - s^0) - ts^1)s^1} = t\frac{x^0}{s^0} + (1-t)\frac{x^1}{s^1} = y^t$. L'élément y^t a donc un antécédent par φ et appartient donc à F. F est donc convexe.

[54] En utilisant les mêmes notations que pour la preuve de la convexité, nous procédons en considérant la contraposé « si un des antécédents de $y \in F$, noté $x \in E$, n'est pas un point extrême de E alors y n'est pas un point extrême de F ». Comme x n'est pas extrême, il existe $x^0, x^1 \in E$ et $u \in]0, 1[$ tel que $x = ux^0 + (1-u)x^1$. En posant $t = \frac{s^0 u}{s^1 + u(s^0 - s^1)}$ on obtient $y = ty^0 + (1-t)y^1$. Or par construction $t \in]0, 1[$ donc y n'est pas un point extrême de F. Par contraposé, on a bien que « si un point de F est extrême alors tous ses antécédents par φ sont des points extrêmes de E ».

Nous nous intéressons maintenant à la condition nécessaire pour que l'état de la cellule soit identifiable par la mesure de V_0, c'est-à-dire à la condition sous laquelle les lieux de V_0 sont disjoints.

Le lieu associé à l'état HH recouvre l'intégralité de l'espace dans lequel V_0 est observable. Ainsi l'état HH peut être confondu avec n'importe quel autre état. Pour pouvoir fonctionner, cette méthode de lecture requiert donc l'interdiction de l'état HH. La même situation se produit pour toutes les valeurs de n avec l'état où tous les éléments mémoires OxRAM sont dans l'état HRS. Nous ignorons donc cet état dans la suite. Pour que les trois états restants soient disjoints, il est nécessaire d'avoir $\frac{h}{l+h} < \frac{l}{1+l}$ et $\frac{1}{l+1} < \frac{l}{h+l}$. Dans les deux cas, cela impose la condition $l^2 > h$.[55]

Pour $n > 2$, nous pourrions procéder de la même manière. Toutefois, alors que la détermination de la condition de séparation des lieux de V_0 pour $n = 2$ pouvait s'obtenir par la transformation d'une simple inégalité, ce ne sera plus le cas pour $n > 2$. En effet, la dimension du problème est toujours $n - 1$. Nous optons donc pour une autre approche de résolution. Nous supposons que la condition trouvée pour $n = 2$ est la même pour $n > 2$ et nous cherchons donc à le prouver. Pour cela, nous allons procéder par récurrence par rapport à n.

Pour commencer, nous effectuons deux remarques :

1. **Symétrie**

   La première porte sur le caractère extrêmement symétrique du problème. La moyenne pondérée définissant V_0 est symétrique par rapport à toute permutation de l'indexation, tant que cette permutation est appliquée globalement, c'est-à-dire sur toutes nos grandeurs à la fois (V_i, a_i, γ_i). Ainsi, les états LH et HL sont symétriques l'un de l'autre par inversion du rôle de a_1 et a_2. De façon générale, tous les états partageant le même nombre de H et de L sont symétriques entre eux.

2. **Restriction au cône**

   La seconde est le fait qu'une situation où un élément mémoire OxRAM a une conductance nulle, est équivalente à celle où l'élément mémoire OxRAM n'est pas présent. Prenons un état E composé de n symboles H ou L. Considérons maintenant l'état EH composé des mêmes n symboles que E avec un H ajouté à la fin. Le lieu de V_0 dans l'état EH restreint à $a_{n+1} = 0$ est identique à celui de E (c'est-à-dire que les a_i sont égaux pour $1 < i < n$ dans les deux cas). Géométriquement cela se traduit par le fait que les lieux de V_0 sur la face du simplexe de dimension n sont identiques aux lieux du simplexe de dimension $n - 1$. De plus, on peut montrer que les lieux correspondants à l'état EH sont contenus dans le cône de sommet V_{n+1} ayant comme base le lieu correspondant de l'état E[56]. On peut en conclure que si les lieux de V_0 pour les différents états possibles avec n éléments mémoires OxRAM sont disjoints, alors tous les états avec $n + 1$ éléments mémoires OxRAM qui se finissent par H le sont également[57]. En combinant ceci avec la symétrie, on peut conclure que si tous les états avec n éléments mémoires OxRAM sont disjoints, alors tous les états avec $n + 1$ éléments mémoires OxRAM ayant au moins un H le sont également.

[55] Remarquons que, tout comme nous l'avons fait avec l'état HH, il serait également possible d'interdire l'état LL. Cela nous laisserait deux états possible qui seraient distinguables si et seulement si $h < l$, c'est-à-dire si et seulement si les états LRS et HRS sont disjoint en conductance absolue. Cette option est celle choisie par la lecture différentielle.

[56] Soit l'état EH avec E une suite de n symbole H ou L et (b_i) un point du lieu de V_0 dans l'état E. Il existe donc (γ_i) tels que $b_i = \frac{\gamma_i}{S}$ avec $S = \sum_{j=0}^{n} \gamma_j$. Considérons ce que devient ce choix de conductance dans l'état EH avec γ_{n+1} quelconque. On a $a_i = \frac{\gamma_i}{\sum_{j=0}^{n+1} \gamma_j} = \frac{\gamma_i}{S+\gamma_{n+1}}$ qui appartient au lieu de V_0 dans l'état EH. Avec $i = n + 1$, on a $a_{n+1} = \frac{\gamma_{n+1}}{S+\gamma_{n+1}}$, et donc $\gamma_{n+1} = \frac{S a_{n+1}}{1-a_{n+1}}$. Pour $i \leq n$, on a $a_i = \frac{b_i S}{S+\gamma_{n+1}}$. En combinant ces deux expressions, on en déduit que pour $i \leq n$, $a_i = b_i(1 - a_{n+1})$. En terme géométrique, cela signifie que le point (a_i) appartient à la droite passant par $\binom{(b_i)}{0}$ et V_{n+1}. Ainsi, le lieu de V_0 dans l'état EH est bien contenu dans le cône de sommet V_{n+1} ayant comme base le lieu de l'état E.

[57] Car l'intersection de deux cônes de même sommet est précisément le cône de même sommet ayant comme base l'intersection des bases des deux cônes. Dans notre cas, l'intersection vide des bases de nos deux cônes implique que l'intersection des deux cônes l'est également.

Avec ces deux observations, nous pouvons presque établir l'hérédité selon n de la relation obtenue dans le cas $n = 2$, sauf pour le cas de l'état composé uniquement de symbole L.

Concentrons-nous donc sur ces états composés uniquement de L. De façon générale, considérons l'état à n élément mémoire OxRAM tout en LRS et un état avec k H et $n - k$ L. Du fait de la symétrie on peut choisir l'ordre des symboles du second état sans perdre en généralités. Ainsi, nous considérons l'état H...HL...L avec k H. Considérons le repère d'origine V_n et d'axe V_nV_i, nous avons alors une correspondance entre les coordonnées des points et le a_i pour $i < n$. Posons $N = (1, ..., 1, 0, ..., 0)^T$ composé de k 1 sur les k premières coordonnées. Notons P le point du premier état le plus dans la direction de $-N$, c'est-à-dire maximisant $-N.P$. De même, notons Q un point du second état maximisant $N.Q$. Finalement, posons $M = \frac{P+Q}{2}$. On souhaite montrer que l'hyperplan de normal N et contenant M est un plan de séparation des lieux des deux états sous notre condition de séparation des lieux de V_0. Pour ce faire, on va montrer que $\overrightarrow{MP}.(-N) < 0$ et $\overrightarrow{MQ}.N < 0$ si et seulement si $h < l^2$. En remplaçant M, on obtient $2\overrightarrow{MP}.(-N) = 2\overrightarrow{MQ}.N = (Q - P).N$, il est donc suffisant de calculer $(Q - P).N$. Par définition, le point P correspond à la situation où les k premières dimensions sont minimales, et les autres maximales, c'est à dire pour $\gamma_i = l$ pour $i \leq k$ et $\gamma_i = 1$ pour $k < i < n$. Ainsi, on a $P = \frac{1}{kl+n-k-1}(l, ..., l, 1, ..., 1)^T$. De même $Q = \frac{1}{kh+l(n-k-1)}(h, ..., h, l, ..., l)^T$. Donc $(Q - P).N = \frac{kh}{kh+l(n-k-1)} - \frac{kl}{kl+n-k-1} = \frac{k(h-l^2)(n-k-1)}{...} < 0$ si et seulement si $h < l^2$. On obtient bien le résultat précédemment énoncé, les états composés uniquement de L sont disjoints des autres états (sauf celui composé uniquement de H) si et seulement si $h < l^2$ pour toutes les valeurs de $n \geq 2$. On peut même en conclure que le point M est le point de rencontre des deux lieux lorsque $h = l^2$ et l'hyperplan de normale N passant par M est un hyperplan de séparation des deux états.

Ainsi nous venons de montrer que pour tout n, les états (hormis celui composé uniquement de H) associé à n éléments mémoires OxRAM sont tous disjoints si et seulement si $h < l^2$. Nous avons également déterminé des hyperplans de séparation entre les états composés que de L et les autres.

En utilisant la règle du cône, nous pouvons en déduire les hyperplans de séparation des autres états. En effet, on peut procéder par récurrence. Considérons l'hyperplan de séparation entre l'état E = L... L et F = H..HL..L pour $n = k$. On en déduit que pour $n = k + 1$, l'hyperplan correspondant à celui de dimension inférieure prolongé sur la nouvelle dimension par l'ajout du point V_k est un hyperplan de séparation entre les états EH et FH. En utilisant la symétrie, cela nous permet d'obtenir un hyperplan de séparation pour toute paire d'états ayant au moins un H au même index.

En conclusion, nous avons montré qu'une cellule 1TnR, $2^n - 1$ états sont séparables en lecture relative à la condition que les conductances des éléments mémoires OxRAM appartiennent à l'intervalle $[0, h\gamma]$ en HRS et à $[l\gamma, \gamma]$ en LRS avec $h < l^2$ et γ quelconque.