Caractérisation électrique et modélisation de la dynamique de commutation résistive dans des mémoires OxRAM à base de HfO2

Clément Nguyen

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Nano-électronique et nano-technologies**

Arrêté ministériel : 25 mai 2016

Présentée par

Clément NGUYEN

Thèse dirigée par **Gérard Ghibaudo**

préparée au sein du **Laboratoire de Caractérisation et de Tests Electriques du CEA-LETI**
dans **l'École Doctorale EEATS de l'INP Grenoble**

Caractérisation électrique et modélisation de la dynamique de commutation résistive dans des mémoires OxRAM à base de HfO_2

Thèse soutenue publiquement le **3 mai 2018**,
devant le jury composé de :

Mme Anne Kaminski
Professeur, Grenoble INP, Présidente du jury
Mr Jean-Michel Portal
Professeur, Université Aix-Marseille, Rapporteur
Mr Damien Deleruyelle
Professeur, INSA Lyon, Rapporteur
Mr Gérard Ghibaudo
Directeur de recherche, CNRS Alpes, Directeur de thèse
Mr Carlo Cagli
Ingénieur de recherche, CEA Grenoble, Encadrant

Résumé

Les mémoires résistives à base d'oxyde OxRAM sont une technologie de mémoire non-volatile dite émergente, au même titre que les mémoires à changement de phase (PCRAM) ou les mémoires magnétorésistives (MRAM). A l'origine les OxRAM étaient très étudiées pour concurrencer les mémoires Flash, dont le fonctionnement est basé sur le stockage de charges dans une grille flottante. Cependant, avec l'avènement des technologies 3D-NAND, il semble très difficile pour les OxRAM d'atteindre les mêmes capacités de stockage que les flashs. Cependant, leur impressionnante vitesse de fonctionnement, bien supérieure à celle des NAND, et leur coût bien inférieur à celui des DRAM, leur permet de se situer à la frontière entre ces deux technologies, dans une catégorie qualifiée de « Storage Class Memory ». De plus, il s'agit d'une technologie dont l'intégration en Back-End-Of-Line, juste au-dessus des circuits CMOS, est très facile, ce qui la rend très attrayante. En revanche, les OxRAM sont connues pour présenter une forte variabilité, et cela représente le principal obstacle à leur démocratisation.

Au cours de cette thèse, nous avons cherché à étudier en profondeur la dynamique de commutation résistive de mémoires OxRAM à base d'oxyde d'hafnium, avec une volonté de se concentrer sur des temps très courts, puisqu'ils représentent l'un des atouts majeurs de cette technologie. Pour cela, ces travaux de thèse se concentrent tout d'abord sur un aspect expérimental, de caractérisation électrique. Nous avons ainsi pu observer, avec un suivi dynamique, la commutation résistive des mémoires, sur des temps de l'ordre de la dizaine de nanoseconde, pour les opérations d'écriture et d'effacement, via la mise au point d'un banc de test entièrement dédié à cette tâche. Ensuite, nous avons analysé les impacts que la réduction du temps de pulse, ainsi que l'abaissement des courants et tensions mis en jeu, peuvent avoir sur la fiabilité des OxRAM, avec des mesures de variabilité. La seconde partie de ce travail de thèse est un travail de modélisation, avec la mise au point d'un modèle physique semi-analytique, dans le but de comprendre les mécanismes de commutation résistives. Après avoir comparé les résultats obtenus par notre modèle aux résultats expérimentaux précédents, nous avons cherché à appliquer notre modèle à des mesures de statistiques. Nous avons ainsi réalisé des tests électriques sur des matrices OxRAM, que nous avons tenté de reproduire avec le modèle. Enfin, nous avons étudié plus en profondeur le bruit à basse fréquence dans les OxRAM, qui constitue l'un des facteurs majeurs de dégradation de la fiabilité des OxRAM, tout en cherchant des pistes pour le diminuer.

Abstract

Oxyde-based resistive memories OxRAM are a technology of emergent non-volatile memory, as phase-change memories (PCRAM) or magnetoresistive memories (MRAM). In the beginning OxRAM were very studied in order to compete with Flash memories, whose mechanism relies on the storage of electrical charges in a flotting gate. However, with the arising of 3D-NAND technology, it seems very difficult for OxRAM to reach the same storage capacities as Flash memories. But their impressive operating speed, far higher than NAND's, and their cost far lower than DRAM's, allow them to operate at the border of these two technologies, in a category called « Storage Class Memory ». Furthermore, the integration of OxRAM in the Back-End-Of-Line, just above CMOS circuits, makes this technology very attractive. On the other hand, OxRAM are known to have a very strong variability, which represents the main obstacle to their expansion.

In this thesis, the dynamics of the resistive switching of hafnium oxyde based OxRAM has been investigated, with a desire to focus on very short times, as they are one of the main assets of this technology. To do so, our work first focuses on an experimental aspect, with electrical characterization. We were able to watch, with a dynamical monitoring, the resistive switching of the memories, at the scale of the dozen of nanoseconds, for writing and erasing operations, thanks to an entirely dedicated set-up. Then, the impacts that the time reduction, and the lowering of the voltage and current, can have on the reliability of OxRAM, were analysed, with variability measurements. The second part of this work concerns modelisation, with the elaboration of a physics-based, semi-analytical model, in order to understand the switching mechanisms. After the comparison of the results obtained by our model with the experimental ones, our model has been applied to statistical measurements. Electrical tests on OxRAM arrays have been performed, and fitted by the model. Finally, the low frequency noise (RTN) in OxRAM has been studied, as it stands as one of the main factors of degradation of OxRAM reliability. Ideas to improve the robustness of OxRAM against RTN are suggested.

Table des matières

Introduction 10

Contexte général 10

Organisation du manuscrit 10

Liste des acronymes 11

Chapitre I : Etat de l'art 14

1. Introduction 14

2. Mémoires conventionnelles 17

2.1. Dynamic Random Access Memory 17

2.2. Static Random Access Memory 17

2.3. Mémoires Flash 18

3. Technologies de mémoires non-volatiles émergentes 20

3.1. Mémoires ferroélectriques FeRAM 20

3.2. Mémoires magnétorésistives STT-RAM 22

3.3. Mémoires à changement de phase PCRAM 23

3.4. Mémoires résistives RRAM 25

4. Mémoires OxRAM 27

4.1. Principes de fonctionnements des OxRAM 27

4.2. Commutation 29

4.3. Matériaux utilisés en tant qu'oxydes 31

4.4. Contrôle du courant de compliance 32

5. Etat de l'art en matière de performances 35

5.1. Miniaturisation 35

5.2. Endurance 37

5.3. Rétention 38

5.4. Consommation en énergie 39

5.5. Vitesse de commutation 40
5.6. Variabilité 43
5.7. Comparaison avec les autres familles de mémoires émergentes 45
6. Modélisation physique 46
6.1. Mécanismes de commutation 46
6.2. Mécanismes de conduction 52
7. Conclusion 54
8. Références du chapitre I 55
Chapitre II : Mesures ultra-rapides et impact sur les distributions 64
1. Introduction 64
2. Protocoles de mesures classiques de dispositifs OxRAM 65
2.1. Mode quasi-statique 65
2.2. Mode pulsé 67
2.3. Mesures d'endurance 69
2.4. Distributions résistives et variabilité 70
3. Présentation du réticule MARS (Mémoire Avancée Résistive à Sélecteur) et des différents splits 71
3.1. Présentation du réticule 72
3.2. Fabrication des dispositifs 73
4. Etude dynamique du switching sur des temps ultra-courts 74
4.1. Présentation du set-up 74
4.2. Confirmation de la fiabilité du véhicule de test 75
4.3. Mesure des capacités parasites 79
4.4. Observation directe du set 85
4.5. Influence de la vitesse de rampe sur la tension de set 87
4.6. Mesure en température 89
4.7. Cas du reset 90
5. Impact sur les distributions 95
5.1. Protocole de mesures 95
5.2. Influence de la tension de set sur les distributions LRS 96
5.3. Résultats en fonction de t_{SET} et Icc sur les distributions LRS 97

5.4. Equivalent pour le reset : influence de t_{RESET} et de Vr 104
6. Conclusion 106
7. Références du chapitre II 107
Chapitre III : Etude de la dynamique de switching via un modèle semi-analytique 110
1. Introduction 110
2. Mise au point d'un modèle de switching 111
2.1. Contexte du modèle 111
2.2. Etude de l'activation de la conduction en température 113
2.3. Etude de la conduction de l'état HRS 116
2.4. Etude de la conduction de l'état LRS 119
2.5. Lien entre la résistance électrique et le nombre de lacunes d'oxygène 120
2.6. Génération de lacunes d'oxygène 124
2.7. Recombinaison de lacunes d'oxygène 125
2.8. Calcul de la température 126
3. Fonctionnement du modèle 127
3.1. Schéma de fonctionnement 127
3.2. Impact de paramètres d'entrée 128
4. Etalonnage du modèle 130
4.1. Détermination de l'énergie d'activation de génération pour HfO_2/Ti 130
4.2. Fit en quasi-statique du set 132
4.3. Opération de reset 133
4.4. Mode discret 135
4.5. Fit en quasi-statique du reset 136
4.6. Tests électriques avec différentes électrodes supérieures 137
4.7. Extraction des paramètres physiques et comparaison avec des simulations *ab initio* 139
5. Conclusion 141
6. Références du chapitre III 141
Chapitre IV : Etude du bruit et de son impact sur la fiabilité 146
1. Introduction 146
2. Présentation des tests sur matrices 147

2.1. Structure du véhicule de test 147
2.2. Présentation du banc de test pour les tests sur matrices 151
2.3. Présentation des tests sur les mémoires 154
3. Fit des distributions 4kbits des tensions de switching par le modèle pour différents temps de pulse 159
4. Impact du forming : fit avec le rayon du filament 161
5. Mesure de bruit à basse fréquence 165
5.1. Etude du bruit à basse fréquence dans la littérature 165
5.2. Set-up expérimental de l'étude du bruit RTN 167
5.3. Analyse du bruit RTN 168
5.4. Impact du courant de forming sur le nombre de niveaux de courant 170
5.5. Impact du courant de forming sur l'amplitude du bruit 172
5.6. Impact du courant de forming sur des lectures successives 176
5.7. Impact du courant de forming sur des lectures de matrices 4kbits 178
6. Conclusion 181
7. Références du chapitre IV 181
Conclusion générale et perspectives 184

Introduction

Contexte général

Ce manuscrit de thèse a pour but de présenter les résultats obtenus après les trois années de thèse passées au sein du Laboratoire de Caractérisation et de Tests Electriques (LCTE) du CEA-LETI à Grenoble. L'objet de cette thèse est les mémoires OxRAM (Oxide-based Resistive Random Access Memory). Les OxRAM appartiennent à la catégorie des technologies de mémoires non-volatiles émergentes, au même titre que les MRAM (mémoires magnétorésistives) ou les PCRAM (mémoires à changement de phase). Ces mémoires reposent sur la commutation de résistance d'un oxyde, et présentent des caractéristiques très prometteuses, notamment en termes de vitesse et de consommation d'énergie. Très étudiées, à l'origine dans le but de concurrencer un jour les mémoires Flash, dont le fonctionnement repose sur le stockage de charges dans une grille flottante, les OxRAM sont aujourd'hui pressenties pour combler le fossé qu'il existe entre les DRAM (Dynamic RAM) et les Flashs. Avec l'avènement des 3D-NAND, il semble en effet très difficile pour les OxRAM d'atteindre les même capacités de stockage que les Flashs. Cependant leur vitesse de fonctionnement bien supérieure à celle des Flashs, à des coûts bien inférieurs aux DRAM leur permet de se situer entre ces deux technologies, dans une catégorie souvent qualifiée de « Storage Class Memory ». Dans ce contexte, ce manuscrit vise à investiguer en profondeur les vitesses que ces mémoires peuvent atteindre, tout en essayant d'optimiser la variabilité de ces dispositifs, souvent décrite comme le principal point faible des OxRAM. Cela passe notamment par la compréhension fine des mécanismes physiques mis en jeu lors des commutations de l'oxyde résistif. C'est la raison pour laquelle, cette thèse est composée à la fois d'une partie « Test Electrique » et d'une partie « Modélisation ».

Organisation du manuscrit

Ce manuscrit sera composé de 4 chapitres. Dans le premier chapitre nous introduirons les différentes technologies de mémoires non-volatiles émergentes. Nous investiguerons en profondeur les différentes performances des OxRAM, dans tous les secteurs d'intérêt, en

s'appuyant sur les nombreux travaux que l'on peut trouver dans la littérature. Enfin, nous les comparerons avec à la fois les technologies déjà matures, et les autres technologies émergentes. Une attention particulière sera portée sur les temps de commutations, puisqu'une grande partie des travaux présentés dans ce manuscrit traitent de ce sujet.

Dans le second chapitre, nous aborderons la partie expérimentale de cette thèse, qui concerne la mise en place et l'optimisation d'un banc de test permettant la mesure, en dynamique, de la commutation des mémoires OxRAM sur des temps ultra-courts. Nous étudierons plus particulièrement l'impact de la vitesse de la rampe de tension utilisée pour commuter l'oxyde, sur les tensions nécessaires à la commutation. Ensuite, nous étudierons l'impact de la minimisation du temps de pulse sur les distributions de résistances des deux états résistifs. Cet impact sera comparé à celui que peut avoir l'abaissement des courants de set et des tensions de reset.

La troisième partie sera consacrée à la mise en place d'un modèle physique semi-analytique destiné à reproduire les courbes de commutation en dynamique des mémoires OxRAM. Nous introduirons tout d'abord les différentes hypothèses et paramètres nécessaires au fonctionnement du modèle. Ensuite, nous nous servirons des mesures réalisées dans le second chapitre pour étalonner le modèle, c'est-à-dire fixer un certain nombre de paramètres. Une fois cela réalisé, le modèle sera considéré comme opérationnel.

Enfin, le quatrième et dernier chapitre aura pour but d'étudier le bruit basse fréquence chez les OxRAM. En effet, ce bruit peut être responsable d'un certain nombre de désagréments lors de l'utilisation d'OxRAM pouvant conduire notamment à des erreurs de lectures. Le modèle détaillé dans le chapitre III sera ici utilisé pour faire le lien entre les résultats expérimentaux et les caractéristiques physiques des OxRAM. Ceci nous conduira au final à proposer des pistes pour abaisser l'impact de ce bruit.

Liste des acronymes

OxRAM : Mémoire résistive à base d'oxyde
RRAM : Mémoire résistive
LRS : Etat faiblement résistif
HRS : Etat hautement résistif
Set : Opération d'écriture (passage de l'état HRS à LRS)
Reset : Opération d'effacement (passage de l'état LRS à HRS)

Read : Opération de lecture

Fenêtre résistive : ratio entre la résistance de l'état HRS et l'état LRS

Courant de compliance : valeur d'intensité de courant que l'on fixe via un appareil externe ou un transistor qui permet de limiter la valeur maximale de courant autorisée à circuler dans le circuit

Chapitre I : Etat de l'art

1. Introduction

La création du premier transistor MOS (Metal Oxide Semiconductor) lança la course à la miniaturisation, qui dicte encore aujourd'hui les mondes de l'industrie et de la recherche dans le domaine de la microélectronique. La fameuse loi de Moore, qui doit son nom à Gordon Moore, cofondateur du géant de la microélectronique, et premier fabricant mondial de semi-conducteur, Intel, fut énoncée en 1965. Cette loi empirique peut se résumer de la façon suivante : le nombre de transistors sur une puce double tous les dix-huit mois. En effet, pour satisfaire la demande croissante des consommateurs de technologies toujours plus performantes, l'industrie microélectronique se doit de réaliser des dispositifs toujours plus petits. Aujourd'hui, cet essor des technologies de l'information peut se ressentir dans tous les domaines de la vie quotidienne. De nos smartphones et ordinateurs portables, en passant par l'apparition des objets connectés, la microélectronique façonne en permanence notre environnement. La quantité d'information que ces dispositifs manipulent augmente chaque année de façon exponentielle. Grâce à ce progrès constant, nous pouvons aujourd'hui surfer sur le net à des débits supérieurs à 1Gbits/s, avec la plus grande simplicité sur nos téléphones portables, ou encore stocker des gigabits de musique sur des appareils de quelques centimètres.

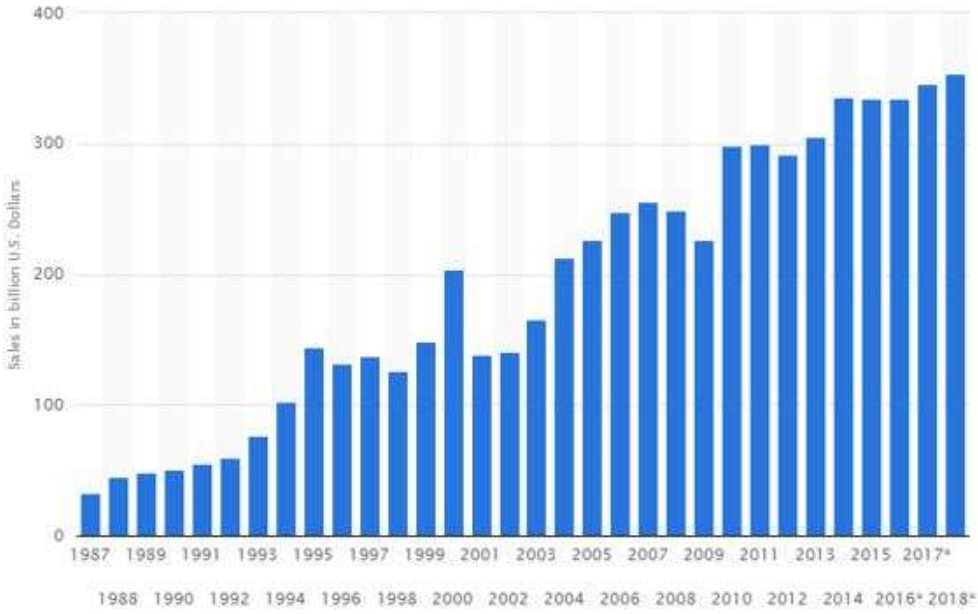


Figure I.1 Evolution du montant des ventes de semiconducteurs de 1987 à 2018 [1].

Ainsi, durant les trente dernières années, le marché des semi-conducteurs a connu une forte croissance, comme en atteste la figure I.1, et approche aujourd'hui les 400 milliards de dollars de ventes.

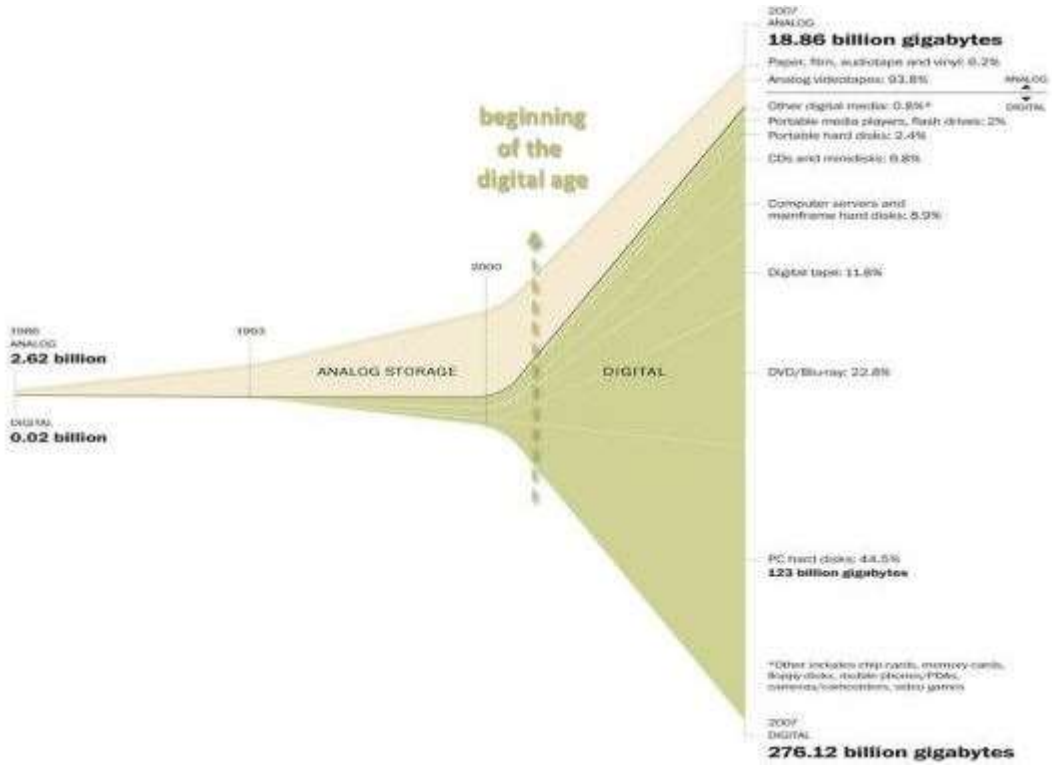


Figure I.2 : Evolution de la quantité de données stockée au cours du temps, selon le type de support [2].

Dans le sillage de cette augmentation des performances, la capacité de stockage des dispositifs augmente également de plus en plus, comme on peut le voir sur la figure I.2. On peut constater que la part de l'information stockée sous forme digitale a explosée peu après les années 2000. Aujourd'hui, près de 94% de l'information est stockée sur support digital (disque dur, CD, mémoires FLASH, etc…). Parmi ces technologies de stockage digitales, la part occupée par les mémoires à base de semiconducteurs est de plus en plus importante. A l'origine dominé par les mémoires volatiles telles que les DRAM (Dynamic Random Access Memory) et les SRAM (Static Random Access Memory), ce marché a commencé à évoluer dans les années 2000 avec l'apparition des mémoires non-volatiles Flash NAND, utilisées notamment dans les clés USB ainsi que les disques durs SSD (Solid State Drive). Aujourd'hui, mémoires volatiles et non-volatiles se partagent assez équitablement le marché (Figure I.3).

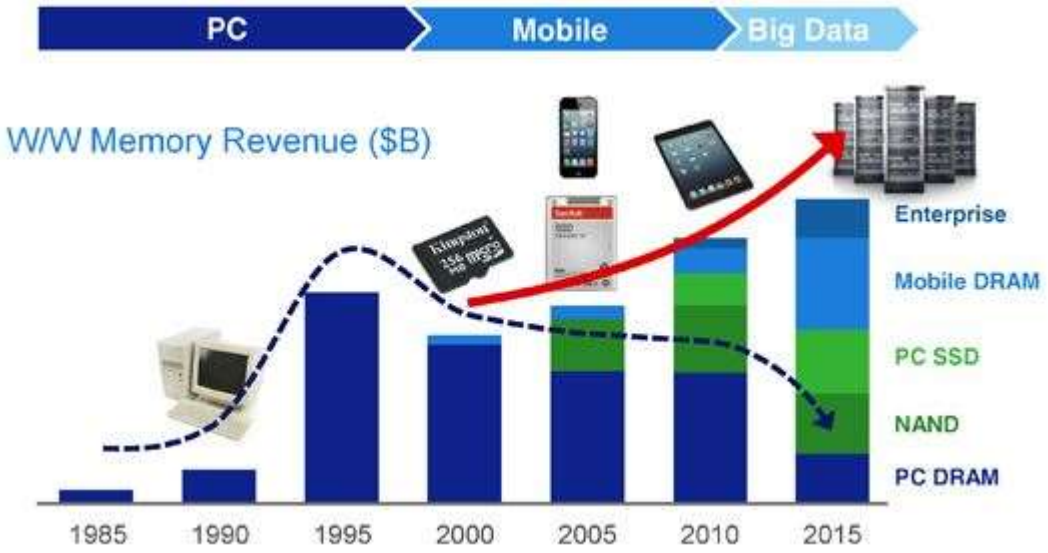


Figure I.3 : Evolution de la part des différentes technologies à base de semiconducteurs entre 1985 et 2015 [3].

La démocratisation de la mémoire Flash NAND s'explique par de bonnes performances en termes de rétention d'information (au-delà de la dizaine d'années) et d'endurance (plus de 10^5 cycles d'écriture/effacement), ainsi qu'à une impressionnante diminution des coûts de fabrication. Ainsi, aujourd'hui, le prix du giga-octet est inférieur à 1$ (cf. Figure I.4).

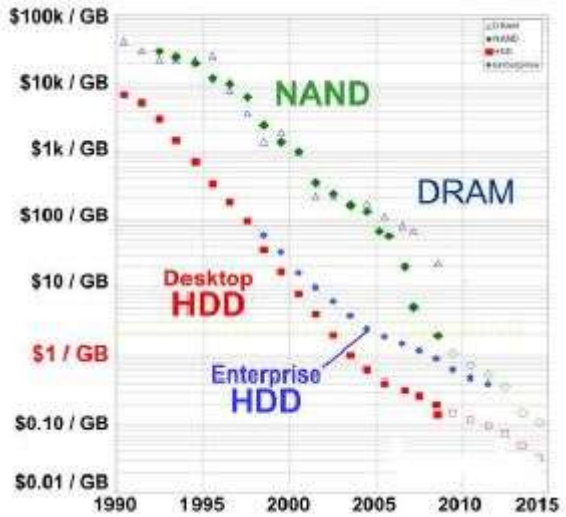


Figure I.4 : Evolution du prix du giga-octet pour différentes technologies [4]

Si au début, la loi de Moore a été respectée avec une étonnante précision, la miniaturisation semble aujourd'hui arriver à bout de course. Les technologies de type Flash semblent en effet atteindre leurs limites. La longueur de grille des transistors atteignant l'ordre de la dizaine de nanomètres, des effets quantiques commencent à apparaître et à menacer la fiabilité de ces dispositifs. Il est ainsi clair que dans les années à venir, les mémoires Flash seront amenées à être remplacés par d'autres dispositifs. Ainsi, d'autres technologies, dites mémoires non-volatiles émergentes (e-NVM) sont actuellement intensément étudiées. Parmi

ces mémoires, on peut citer les MRAM (Magnetoresistive RAM), les FeRAM (Ferroelectric RAM), les PCRAM (Phase-Change RAM) ou encore les RRAM (Resistive RAM).

2. Mémoires conventionnelles

Comme nous l'avons dit plus haut, il y a deux principales familles de mémoires : les mémoires volatiles, qui perdent l'information stockée lorsque l'alimentation est coupée, et les mémoires non-volatiles, qui n'ont pas besoin d'alimentation pour conserver cette information. Ainsi, les SRAM et DRAM appartiennent à la première catégorie, tandis que les mémoires Flash font partie de la seconde. Dans cette partie, nous expliquerons les principes de bases de fonctionnement de ces différentes technologies, et nous indiquerons quelles sont leurs points faibles et points forts.

2.1. Dynamic Random Access Memory

Les DRAM sont d'une technologie assez peu onéreuse. Elles sont principalement utilisées en tant que RAM dans les ordinateurs. Une mémoire DRAM est constituée d'un transistor MOS, ainsi que d'un pico-condensateur (une capacité MIM, Métal/Isolant/Métal). Cette simplicité structurelle en fait un composant facilement miniaturisable et permet d'obtenir des densités importantes. Le bit mémoire est codé par le fait que le condensateur soit chargé ou non : l'état « 1 » est codé par la présence de charges aux bornes du condensateur, tandis que l'état « 0 » est codé par l'absence de charges. Afin de maintenir cette charge dans le condensateur, la mémoire a besoin d'être rafraichie à des périodes de l'ordre de la milliseconde. Il s'agit donc d'une technologie assez gourmande en énergie. Il s'agit cependant d'une mémoire assez rapide, avec des temps d'écriture/effacement de l'ordre de 5 nanosecondes, et très endurante, puisqu'elle peut supporter un nombre presque illimité de cycles (10^{15} cycles) [5,6].

2.2. Static Random Access Memory

Une mémoire SRAM est une mémoire volatile constituée en général de 6 transistors (pour un bit). Cela en fait une technologie bien plus onéreuse et moins facilement miniaturisable que les DRAM. En revanche, elle n'a pas besoin d'être rafraichie périodiquement. De même que les DRAM, les SRAM sont extrêmement endurantes, et encore plus rapides (temps

d'écriture/effacement de l'ordre d'une nanoseconde) [5]. Elles sont en général utilisées en tant que mémoire cache ou tampon.

2.3. Mémoires Flash

Les mémoires Flash appartiennent, contrairement aux SRAM et DRAM, à la catégorie des mémoires non-volatiles. Comme elle l'est schématisée en figure I.5, une mémoire flash se base sur un transistor MOSFET, dans lequel est insérée une grille flottante, entre la grille et le substrat. Cette grille flottante est isolée du substrat par un oxyde tunnel, et de la grille de contrôle par un oxyde dit de contrôle. Ici le codage de l'information se fait par le piégeage ou le dépiégeage de charges dans la grille flottante [7]. En effet, la présence de charge dans la grille flottante modifie la tension de la grille flottante et augmente donc la tension de seuil. Ainsi, la présence de charge dans la grille flottante code pour un « 0 » tandis que l'absence de charge code pour un « 1 ».

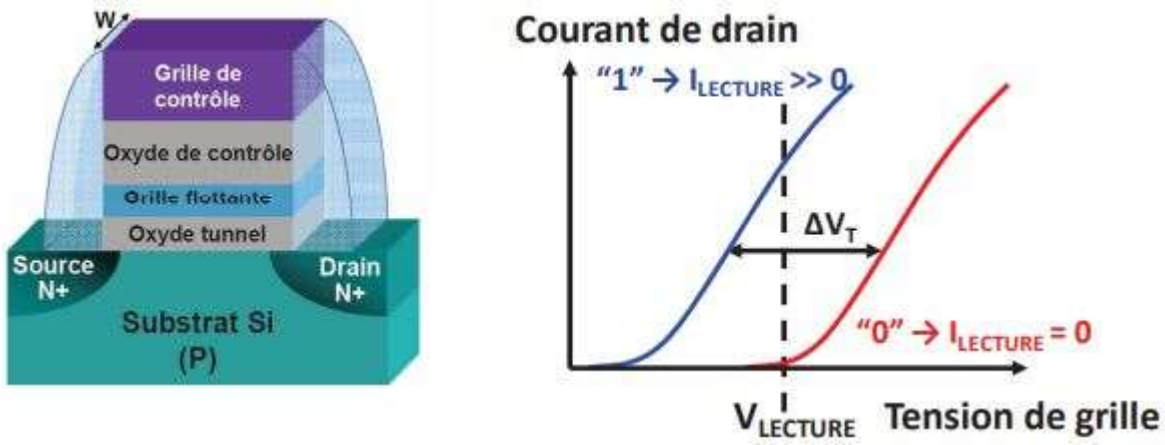


Figure I.5 : (gauche) Schéma d'une mémoire Flash. (droite) Exemple de caractéristique I-V [8]

En appliquant une tension positive suffisamment élevée sur la grille de contrôle, on injecte, par effet Fowler-Nordheim ou par effets d'électrons chauds en fonction de la technologie employée, des électrons dans la grille flottante, à travers l'oxyde tunnel. De même, l'application d'une tension négative permet de libérer des électrons piégés dans cette grille flottante, pour les renvoyer vers le canal.

Il existe deux grandes familles d'architecture pour les mémoires Flash : l'architecture NOR et l'architecture NAND (cf. Fig I.6) :

- Dans l'architecture NOR, les transistors sont connectés en parallèle, ce qui permet un accès individuel de chaque cellule mémoire. Cependant, cette configuration ne permet pas d'obtenir des densités importantes, c'est pourquoi, l'architecture est surtout utilisée pour stocker les codes, dans des applications embarquées [10].

- Dans l'architecture NAND, les transistors sont connectés en série. Ainsi le drain de chaque cellule n'est plus accessible de façon individuelle, mais de façon séquentielle. En revanche, elle permet d'obtenir des densités bien plus importantes, et des coûts bien moindres. Il s'agit donc d'une configuration employée pour stocker des quantités importantes de données [9]. Les NAND flash sont également utilisées dans les clés USB ou les cartes SD.

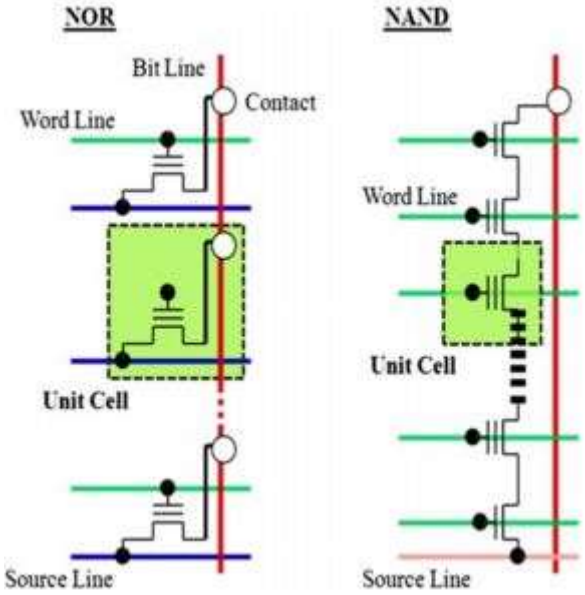


Figure I.6: Architectures NOR et NAND [9]

Les performances typiques des mémoires Flash dépendent de l'architecture employée. Les tensions de programmation sont en général de l'ordre de la quinzaine de Volts [11]. Les temps d'écriture/effacement sont compris entre 10µs et 1ms, tandis que les temps de lecture sont de 50ns pour l'architecture NOR, et de 10µs pour les NAND [12]. D'un point de vue de l'endurance, les mémoires Flash peuvent supporter environ 10^5 cycles.

Ces dernières années beaucoup de travaux ont été réalisés sur les mémoires Flash, qui sont actuellement industrialisées sous la forme de 3D NAND, ce qui a permis d'aller plus loin dans les performances de miniaturisations des mémoires Flash [13].

Les mémoires Flash sont aujourd'hui présentes dans tous les dispositifs microélectroniques modernes, tels que nos smartphones ou nos clés USB. Cependant, à mesure que les dimensions des composants se doivent de diminuer, des effets quantiques font leur apparition et commencent à diminuer la fiabilité des mémoires Flash. En fonctionnement normal, les électrons piégés dans la grille flottante sont censés y rester, malgré la mise hors tension de la mémoire. Cependant, en diminuant les dimensions de la mémoire, et notamment l'épaisseur de l'oxyde tunnel, les charges deviennent plus à même de franchir cette barrière par effet tunnel [14]. A l'avenir, ce problème va limiter la miniaturisation des mémoires Flash. De

plus, cette augmentation de la densité expose les mémoires à des interférences dues à des phénomènes de couplage entre différentes cellules.

C'est pourquoi de nouvelles technologies de mémoires émergentes voient le jour. En plus d'assurer la relève de la technologie flash, ces mémoires offrent des performances proches des SRAM en termes de vitesse, et proches des DRAM en termes de densités d'intégration, tout en conservant le caractère non-volatile des mémoires Flash.

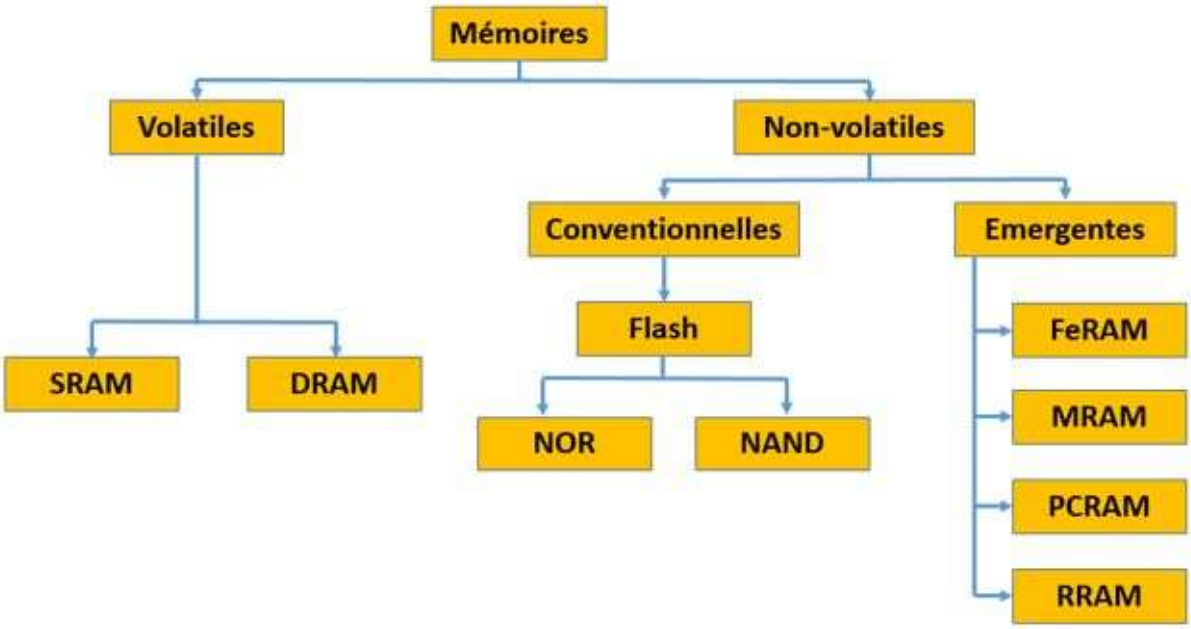


Figure I.7 : Schématisation des différentes classes de mémoires [15]

3. Technologies de mémoires non-volatiles émergentes

Dans cette partie nous ferons un résumé des différentes technologies de mémoires émergentes. Nous introduirons donc les FeRAM, les MRAM, les PCRAM, ainsi que les différentes catégories de RRAM (CbRAM et OxRAM).

3.1. Mémoires ferroélectriques FeRAM

Les mémoires FeRAM font partie des premières mémoires émergentes à avoir été transférées en production. Elles sont constituées le plus souvent de plomb, zirconium et titane (PZT) [9, 16]. Il s'agit d'un matériau ferroélectrique, dans lequel un dipôle électrique existe, même en l'absence de champ électrique extérieur. En appliquant un champ électrique, on peut switcher la mémoire d'un état de polarisation à un autre (cf. Fig I.8), en changeant la position

d'atomes de zirconium et de titane. Ce changement de position subsiste lorsque le champ électrique disparaît, ce qui confère à ce dispositif son caractère non-volatile.

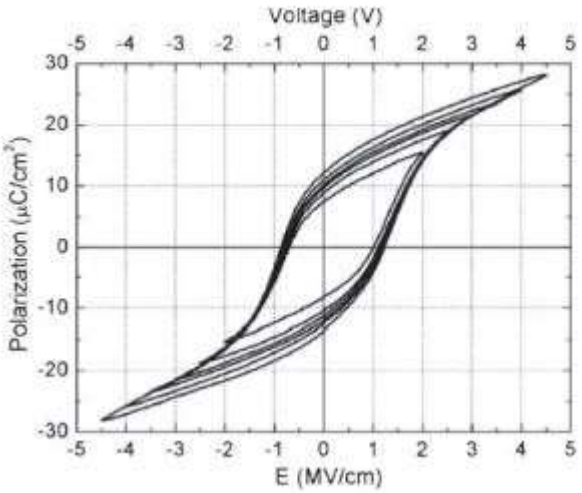


Figure I.8 : courbe d'hystérésis obtenue sur des dispositifs d'oxyde d'hafnium dopés au silicium [17].

L'architecture d'une FeRAM est très proche de celle d'une DRAM : le diélectrique utilisé chez les DRAM est remplacé par ce matériau ferroélectrique. Celui est ainsi associé à un transistor MOS (cf. Fig. I.9).

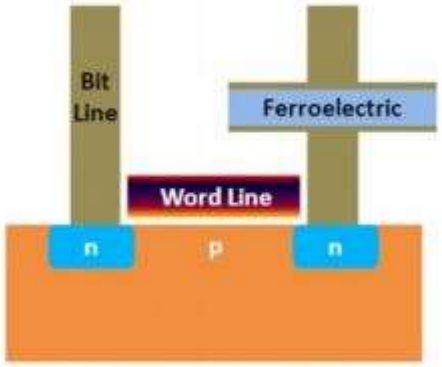


Figure I.9 Schématisation d'une mémoire FeRAM [9]

La lecture de l'état de la mémoire se fait en réalisant une opération d'écriture : si on détecte un pulse de courant, c'est que la mémoire était dans l'état OFF [9]. Il s'agit de l'un des inconvénients de ces mémoires ; l'opération de lecture est destructive. En revanche, les FeRAM sont très rapides, consomment très peu d'énergie, fonctionnent à faible tension (environ 2V) et peuvent supporter un très grand nombre de cycles ($>10^{12}$) [17].

On retrouve ce mécanisme d'opération dans les dispositifs FeFET (Ferroelectric Field Effect Transistor). Ces transistors ferroélectriques non-volatiles sont actuellement très étudiés car ils présenteraient des vitesses supérieures aux DRAM et une meilleure densité que les mémoires Flash [18].

3.2. Mémoires magnétorésistives STT-RAM

Le transfert de spin (STT, Spin Transfer Torque) est un phénomène observé lorsqu'un courant polarisé en spin traverse un matériau magnétique. Lorsque le matériau utilisé est suffisamment fin, on peut observer un transfert du spin du courant polarisé vers ce matériau. On peut donc agir sur l'aimantation de cette nanocouche de matériau, sans utiliser de champ magnétique externe.

Le composant de base d'un MRAM est une jonction tunnel magnétique (MTJ) [19, 20, 21, 22]. Il s'agit d'une structure épaisse de quelques dizaines de nanomètres constituée de trois couches : deux couches magnétiques séparées par une couche d'oxyde. L'une des couches magnétiques, que l'on nomme couche de référence, possède une aimantation stable, qui va servir de référence tout au long de l'utilisation de la MTJ. L'autre couche, appelée couche libre est programmable : son aimantation sera modifiable, lors des phases d'écriture et d'effacement.

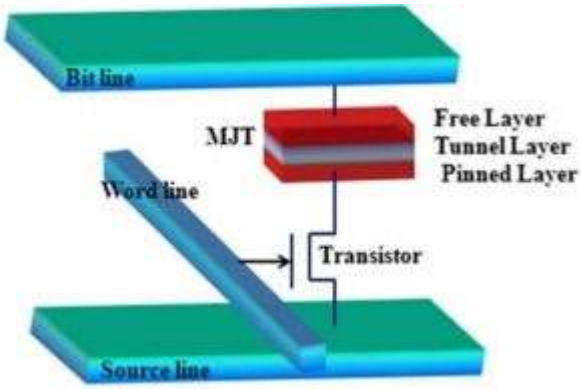


Figure I.10 : Schéma d'une mémoire STT-RRAM [9]

L'opération de lecture se base sur le phénomène de magnétorésistance tunnel [17]. Lorsque les aimantations des deux couches magnétiques sont parallèles, la résistance de la jonction tunnel sera basse, tandis que lorsque les deux aimantations sont en configuration antiparallèles, la résistance sera élevée (cf. Figure I.11). Ainsi, le bit mémoire est codé par la résistance de la cellule.

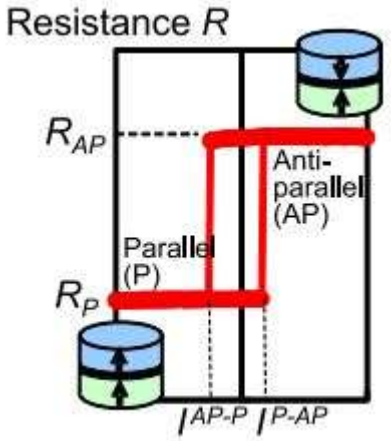


Figure I.11 : Lecture de la résistance d'une MRAM dans les états parallèles et antiparallèles [19]

Les opérations d'écriture et d'effacement consistent à faire switcher l'aimantation de la couche libre, en utilisant le principe de transfert de spin, décrit ci-dessus.

En termes de performances, les STT-MRAM sont extrêmement rapides [23], fonctionnent à faible tension et disposent d'une endurance quasi-illimitée. En revanche, elles présentent une fenêtre résistive (c'est-à-dire le ratio entre les résistances des états ON et OFF) assez faible [16]. De plus les états résistifs ne sont pas très stables thermiquement et les matériaux utilisés ne sont pas parfaitement compatibles avec les procédés de fabrication CMOS [24].

3.3. Mémoires à changement de phase PCRAM

De même que les STT-MRAM, les PCRAM font partie de la famille des mémoires à changement de résistance [25] : le bit mémoire est codé par la résistance de la cellule. Comme leur nom l'indique le changement de résistance des PCRAM est lié à un changement de phase au sein du dispositif.

Une mémoire PCRAM est constituée de deux couches (un isolant et un matériau à changement de phase, très souvent du GST (un alliage de germanium, d'antimoine et de tellure)) prises en sandwich entre deux électrodes (cf. Figure I.12a). Le GST, qui appartient à la famille des chalcogénures. Un tel matériau peut exister sous deux formes : une forme cristalline et une forme amorphe. Ces deux phases présentent des caractéristiques différentes, notamment en termes de résistivité [26, 27] : l'état cristallin présente une faible résistivité, tandis que la résistivité de l'état amorphe est supérieure de plusieurs ordres de grandeurs. En chauffant ce matériau il est possible de le faire changer d'état. C'est pourquoi, un matériau

conducteur (appelé heater) relie l'électrode et le GST, à travers l'isolant, afin de chauffer une partie du GST (Figure I.12a).

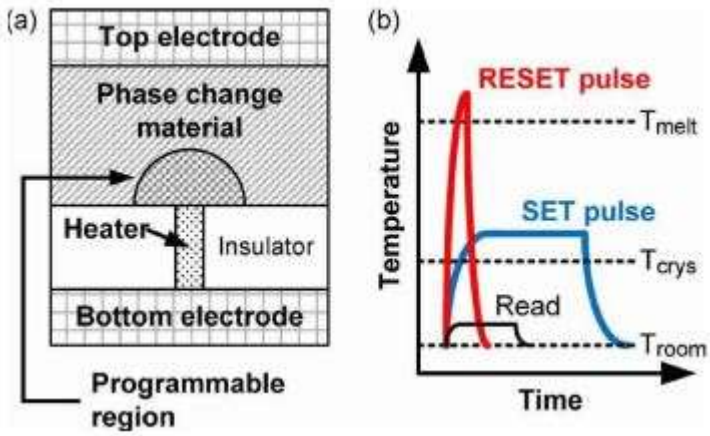


Figure I.12 : (a) Schéma d'une mémoire PCRAM. (b) Programmation et lecture d'une PCRAM via l'application d'un champ électrique [28].

Le cyclage d'une PCRAM s'effectue de la façon suivante : après fabrication, le GST est dans l'état cristallin. Pour réaliser un reset, une partie du GST est chauffée via l'application d'un courant électrique relativement élevé. Cette partie va alors fondre et passer à l'état amorphe, de haute résistivité. Pour retourner à l'état faiblement résistif, un courant électrique plus faible est appliqué et va chauffer le matériau à une température située entre la température de cristallisation et la température de fusion (Fig. I.12(b)), sur une durée suffisamment longue pour permettre la cristallisation du matériau.

Un travail important a été réalisé afin d'abaisser le courant de reset, nécessaire à la fusion du matériau, notamment en réduisant les dimensions de la cellule mémoire (Fig. I.13). Cependant, la consommation en courant, reste l'un des principaux défauts de cette technologie. Malgré cela, les mémoires PCRAM sont souvent considérées comme l'une des technologies de mémoires non-volatiles émergentes les plus matures, de par notamment leur grande vitesse, leur grande endurance et la possibilité de fabrication à grande échelle [26].

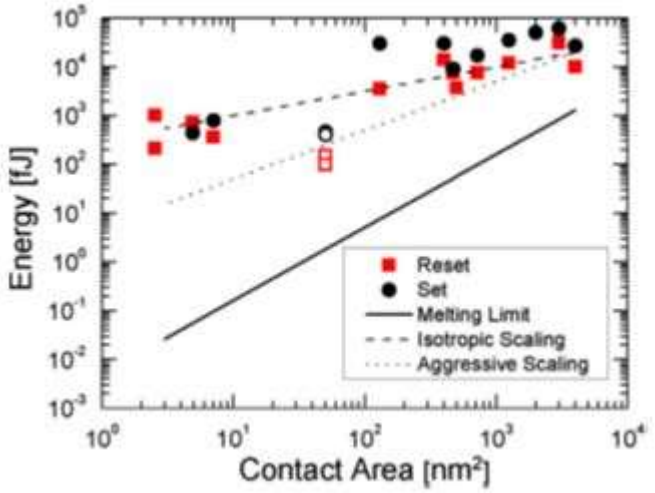


Figure I.13 : Evolution de la consommation en énergie en fonction de la surface du point mémoire [29].

3.4. Mémoires résistives RRAM

La dernière classe de mémoires non-volatiles émergentes est celle des mémoires résistives (ReRAM). Celles-ci peuvent être divisées en deux sous-catégories : les CbRAM (Conductive Bridge RAM) et les OxRAM (Oxide-based RAM).

a) CbRAM

Une CbRAM est composée d'un électrolyte solide, prise en sandwich entre deux électrodes métalliques (cf. Fig. I.14). L'une de ces électrodes (généralement en cuivre ou en argent), servira d'électrode active. En effet, comme chez les PCRAM, le bit mémoire est codé par la résistance de la cellule [30, 31]. Le changement de résistance se fait via la création ou la dissolution d'un filament conducteur à travers l'électrolyte, qui relie les deux électrodes. Ce filament est formé par les ions métalliques issus de l'électrode active. L'écriture consiste en la création de ce filament, tandis que l'effacement est l'opération de destruction d'une partie de ce filament, via la re-migration de ces ions métalliques vers l'électrode active.

Ces opérations sont réalisées via l'application d'une tension électrique entre les deux électrodes. En effet, une tension positive appliquée sur l'électrode supérieure va entraîner la migration des cations métallique vers la zone du filament, tandis qu'une tension opposée va faire migrer les cations du filament vers l'électrode active [32, 33]. Ainsi, c'est la migration des ions métalliques dans l'électrolyte qui permet le fonctionnement de ces mémoires.

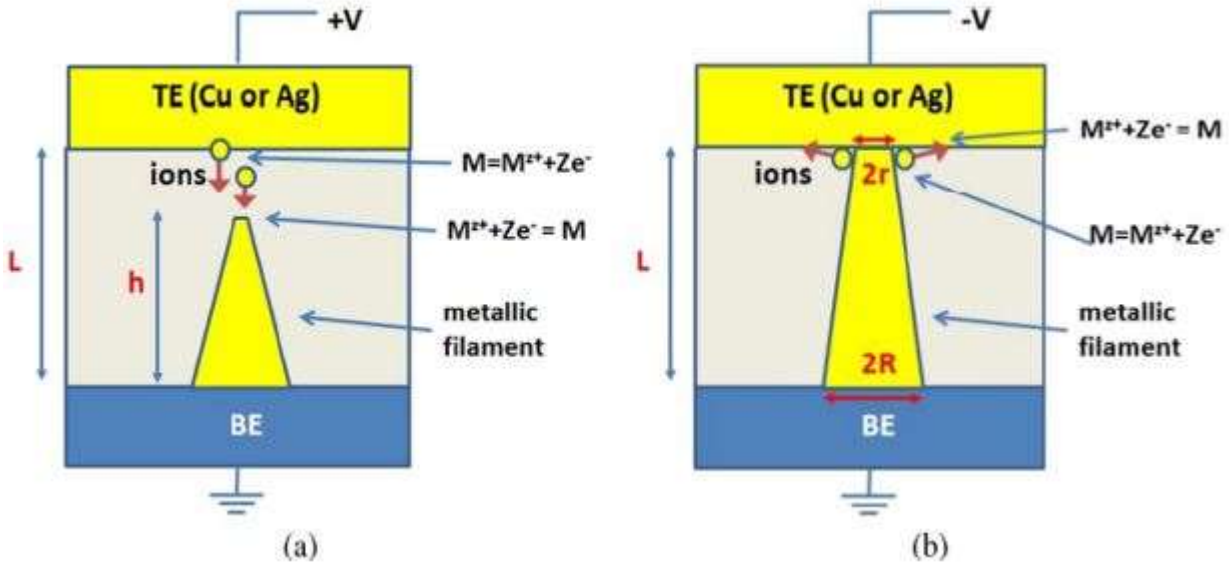


Figure I.14 : Schématisation du fonctionnement d'une CbRAM [30]

Les matériaux qui constituent l'électrolyte solide sont en général des chalcogénures (comme GeSe ou GeS) [32, 34] ou des oxydes (de tungstène [34], de silicium [33], etc…).

Les CbRAM présentent des performances très compétitives, notamment en termes de vitesse [35], d'endurance et de fenêtre résistive [16]. En revanche, les CbRAM souffrent d'une relative mauvaise rétention, et d'instabilités en température [16, 36].

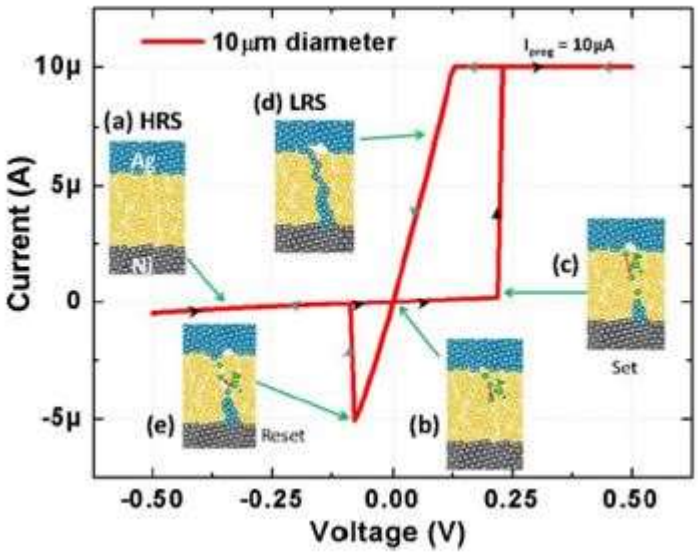


Figure I.15 : Exemple de courbes courant-tension d'opérations de set et reset [32]

a) OxRAM

Les OxRAM sont relativement proches des CbRAM en terme de fonctionnement. En effet, le mécanisme de commutation est également d'origine filamentaire. La différence est liée à la nature du filament conducteur. Comme leur nom l'indique, les OxRAM sont basées sur des oxydes métalliques (HfO_2, NiO ou Ta_2O_5, par exemple). La structure d'une OxRAM est très

basique : la couche d'oxyde métallique est simplement entourée par deux électrodes (dans une simple structure MIM) métalliques (cf. Fig. I.16).

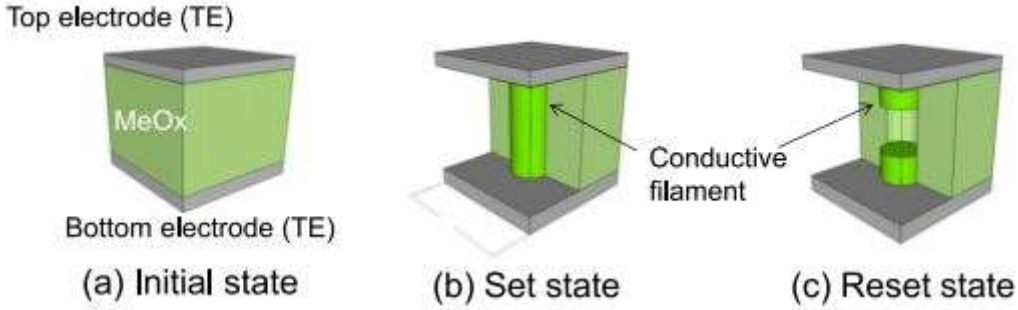


Figure I.16 : Schéma simplifié du fonctionnement d'une OxRAM, basé sur un oxyde métallique [37]

Alors que chez les CbRAM, le filament est constitué d'atomes ou d'ions métalliques issus de l'électrode, il est ici, constitué de lacunes d'oxygènes : chez les OxRAM, c'est la migration des ions oxygène qui permet le switching de la mémoire [38, 39]. L'opération de set consiste en la création de ce filament, tandis que le reset consiste en la destruction d'une portion de ce filament.

Dans la mesure où ce manuscrit traite des mémoires OxRAM, nous allons étudier plus en détail dans la partie suivante le fonctionnement ainsi que l'état de l'art concernant cette famille de mémoires non-volatile.

4. Mémoires OxRAM

4.1. Principes de fonctionnements des OxRAM

a) Forming

Comme on peut le voir sur la figure I.16, dans son état initial, la mémoire ne possède pas de filament conducteur. Elle est dans un état dit vierge. Pour créer un filament, il faut réaliser au préalable une opération de forming (cf. Fig. I.17). Cette opération consiste à appliquer une tension relativement élevée (entre 2 et 5V, en fonction des matériaux utilisés) sur la top électrode. Cette tension va entraîner la génération de lacunes d'oxygène (via une forme de « soft breakdown » du diélectrique), et la création d'un filament conducteur reliant les deux électrodes. On passe alors d'un état très fortement résistif (plusieurs GΩ) à un état LRS, faiblement résistif (entre 10^3 et 10^4Ω, en général) [40, 41, 42]. A noter que l'état extrêmement résistif pré-forming, ne pourra pas être retrouvé. Le forming est une opération non-réversible :

la résistance de l'état HRS, obtenu par l'opération de reset sera un état intermédiaire entre l'état vierge et l'état LRS.

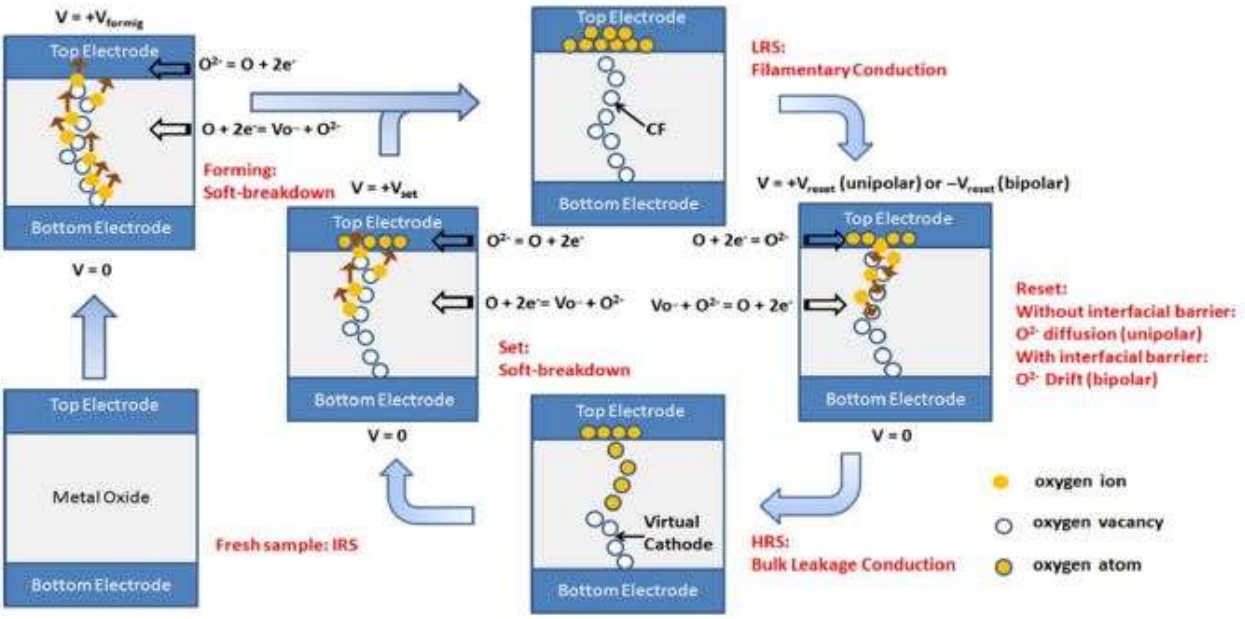


Figure I.17 : Mécanisme de switching lors des étapes de forming, reset et set [40]

Comme précisé ci-dessus, l'oxyde utilisé pour les OxRAM est un oxyde métallique. Lorsqu'une tension positive est appliquée sur la top électrode, des ions d'oxygène vont être arrachés du réseau et vont migrer vers la top électrode sous l'action du champ électrique, puisqu'ils sont chargés négativement. Cela va créer un filament conducteur, constitué de lacunes d'oxygène. Nous reviendrons plus tard, sur les mécanismes physiques mis en jeu.

b) Reset

En appliquant une tension suffisamment élevée, via l'opération de RESET (dont les ordres de grandeurs seront évalués ensuite), il est possible de supprimer une partie du filament conducteur : les ions d'oxygène vont venir combler une partie des lacunes d'oxygène (cf. Fig. I.17). Ainsi la cellule va prendre une valeur de résistance plus élevée. Cependant, le fil ne sera pas totalement détruit, c'est pourquoi la résistance reste bien inférieure à sa valeur avant forming.

c) Set

L'opération de SET est similaire à celle de forming et a pour but de reformer le filament, partiellement rompu suite à l'opération de RESET (cf. Fig. I.17). Dans la mesure où le filament n'est pas totalement détruit, la tension requise pour l'opération de SET est bien inférieure à celle de forming (de l'ordre de 1V). A noter que contrairement à la tension de forming qui dépend de l'épaisseur de la couche, ce n'est plus le cas de la tension de set.

4.2. Commutation

En fonction des matériaux utilisés comme oxyde métallique [40], ou comme électrode [43], différents types de commutation peuvent survenir. Si la commutation d'un état à un autre se fait toujours via des valeurs seuils de tension, la polarité selon laquelle on applique ces tensions est importante et peut varier d'un dispositif à un autre. Alors que le passage de l'état hautement résistif à l'état faiblement résistif (opération de set et forming) se fait toujours via une tension positive appliquée sur la top électrode, il y a deux types de comportements différents concernant le passage de l'état faiblement résistif à l'état hautement résistif (opération de RESET):

- Comportement unipolaire (cf. Fig. I.18(a)): la polarisation utilisée pour le RESET est la même que pour le SET. Dans ce cas, la diffusion des ions d'oxygène (qui viennent combler les lacunes constituant le fil conducteur) est activée par le chauffage par effet Joule lié au passage du courant [44, 45, 46, 47].

- Comportement bipolaire (cf. Fig. I.18(a)): la polarisation utilisée lors du RESET est opposée à celle utilisée pour le SET : pour le RESET, on peut soit polariser négativement la top électrode, soit positivement la bottom électrode. Dans ce cas, c'est surtout la tension appliquée qui commande la migration des ions d'oxygène : celle-ci est dirigée par le champ électrique appliqué [44, 48, 49, 50, 51].

- Comportement non-polaire (Fig. I.19) : cas plus marginal où la polarisation n'importe pas. En effet, on peut réaliser un RESET peu importe la polarisation utilisée lors des opérations de SET et RESET [43].

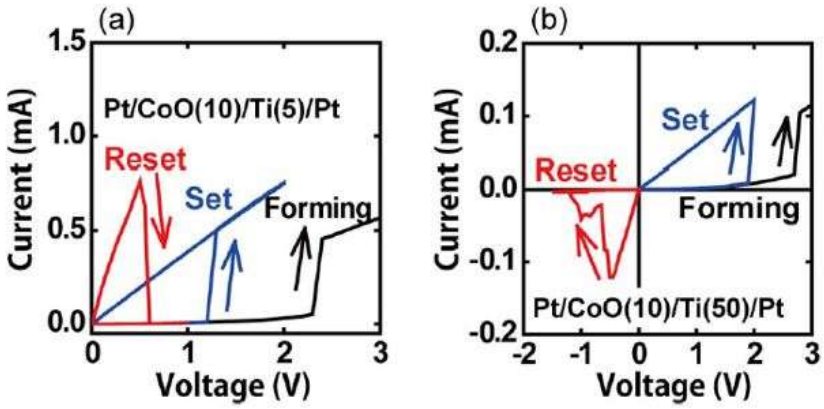


Figure I.18 : représentation des sauts de résistance liés au forming (diminution de la résistance à forte tension), au SET (diminution de la résistance à faible tension) et au RESET (augmentation de résistance). a) Commutation unipolaire b) Commutation bipolaire [42]

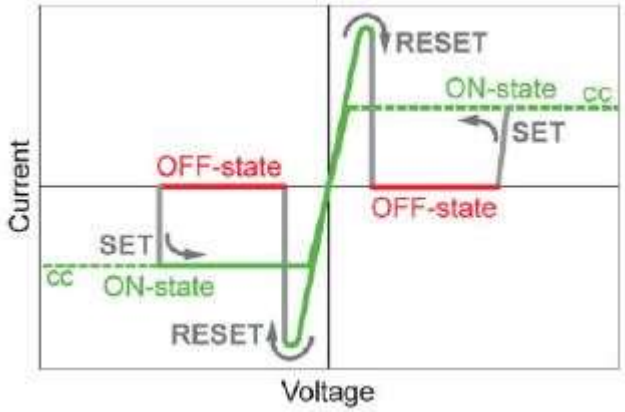


Figure I.19 : Fonctionnement non-polaire [52]

Il est important de noter que le mode de switching ne dépend pas uniquement de l'oxyde utilisé [53, 54]. En effet, on peut constater, par exemple que le choix de deux électrodes inertes (comme en platine) favorise un comportement de type unipolaire, alors que prendre une électrode supérieure active (c'est-à-dire pouvant s'oxyder, comme le titane ou l'hafnium) favorise un comportement bipolaire [43, 54, 55]. En figure I.20, on peut voir un tableau récapitulant le mode de switching en fonction de différents matériaux utilisés.

Unipolar	Bipolar
Pt/NiO/Pt [7]	Pt/NiO/$SrRuO_3$ [54]
Pt/TiO_2/Pt [8]	Pt/TiO_2/TiN [55]
Pt/ZnO/Pt [18]	TiN/ZnO/Pt [56]
Pt/ZrO_2/Pt [57]	Ti/ZrO_2/Pt [57]
Pt/HfO_2/Pt [58]	TiN/HfO_2/Pt [59]
Pt/Al_2O_3/Ru [60]	Ti/Al_2O_3/Pt [61]

Figure I.20 : Influence des matériaux utilisés sur le mode de commutation [40] (les références indiquées correspondent à celle du papier d'où provient le tableau)

Le choix des électrodes ne conditionne pas uniquement le mode de commutation. En effet, une électrode active, telle que le titane, permet également d'abaisser très fortement les tensions nécessaires au switching résistif et au forming [42, 43, 55]. Comme l'électrode est oxydable, les ions oxygène réagissent avec celle-ci. Il en résulte la formation d'une zone située à l'interface entre l'électrode active et l'oxyde métallique, particulièrement riche en lacune d'oxygène (cf. Fig. I.21). C'est la formation de cette zone intermédiaire qui permet l'abaissement des tensions de set et de forming [55, 56]. Au contraire, l'électrode se comporte ensuite comme un réservoir d'oxygène pour les opérations de reset.

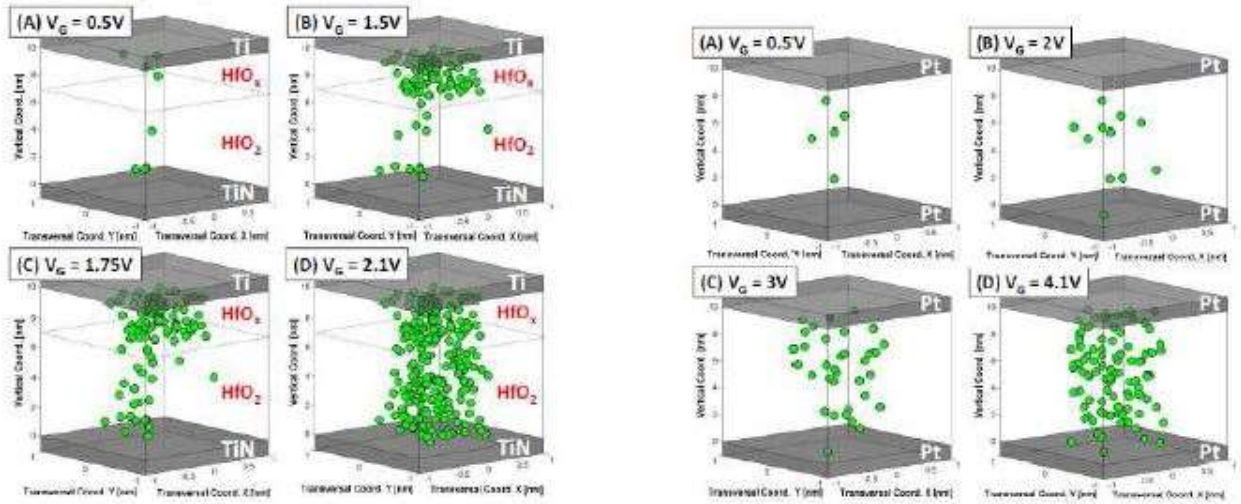


Figure I.21 : simulation de présences de lacunes d'oxygène (en vert) avec des électrodes de Ti/TiN et de Pt. On observe la présence d'une couche intermédiaire riche en lacune, avec Ti. Le filament se forme pour des tensions plus faibles. [55]

4.3. Matériaux utilisés en tant qu'oxydes

Un grand nombre d'oxydes métalliques sont connus pour présenter un comportement de commutation résistive. L'oxyde de nickel NiO est l'un des tout premiers oxydes à avoir été étudiés dans la littérature [57, 58]. Présentant un switching unipolaire et des performances intéressantes [59, 60, 61] (cf. Fig. I.22), les OxRAM basées sur NiO ont une assez mauvaise uniformité, ce qui les rend peu adaptées au monde industriel.

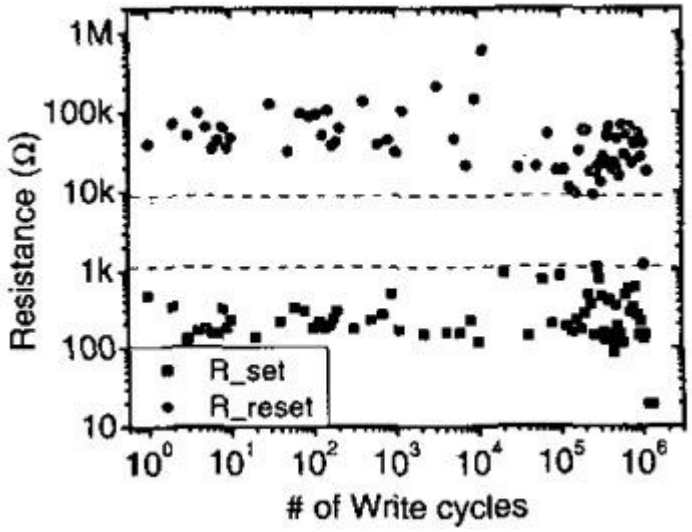


Figure I.22 : Endurance d'un million de cycle obtenue avec des dispositifs de NiO [59]

Aujourd'hui, les mémoires à base d'oxyde de hafnium sont, d'assez loin, les mémoires les plus étudiés, à la fois dans la littérature académique, que dans des papiers à vue plus industrielles [40, 62-66] (cf. Fig. I.23). En effet, souvent associée à une électrode active (comme en Ti ou Hf), les mémoires à base d'oxyde de hafnium présentent d'excellentes caractéristiques en termes de tension [64], de temps [67], d'endurance [64], de miniaturisation [67, 68] ou de

rétention [69, 70]. De plus il s'agit d'un matériau très facilement compatible en BEOL (Back End Of Line) en technologie CMOS [71].

Enfin, même si beaucoup d'autres matériaux sont étudiés dans la littérature (comme TiO_x [73, 74] ou encore AlO_x [75, 76]), les dispositifs les plus évoqués, après ceux à base d'oxyde d'hafnium sont ceux à base d'oxyde de tantale, de par notamment d'excellentes performances en termes de stabilité thermique [77], d'endurance [78], avec de très basse tension d'application [79]. Ainsi, des microcontrôleurs commerciaux de chez Panasonic se basent sur des RRAM à base de TaOx [81] (cf. Fig. I. 24).

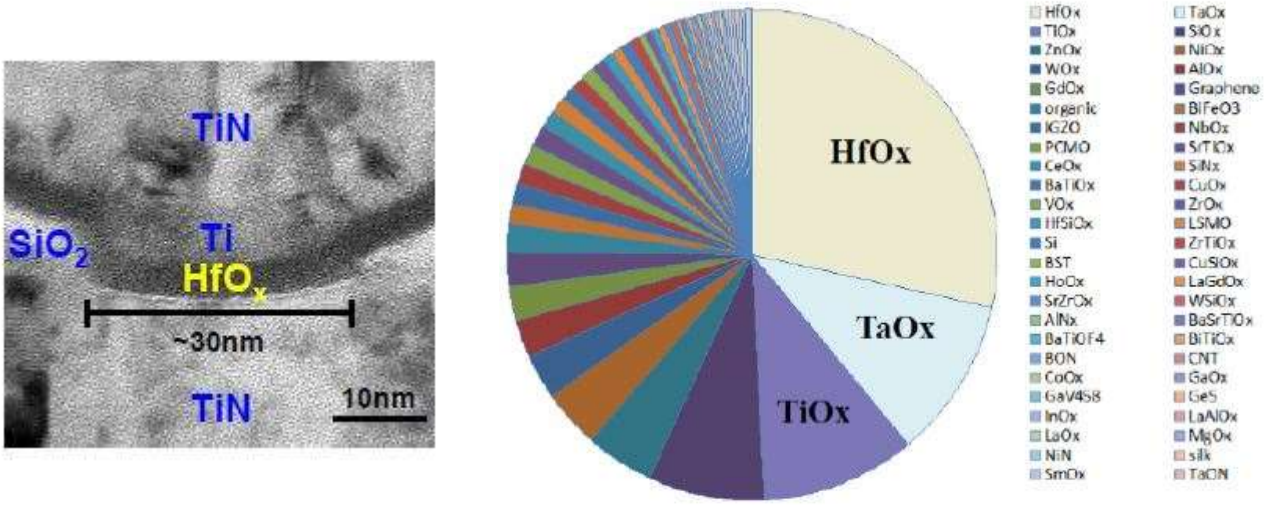


Figure I.23 : (gauche) image XTEM d'un empilement TiN/TiOx/HfOx/TiN [72]; (droite) différents oxydes utilisés dans les publications de l'année 2013 [62]

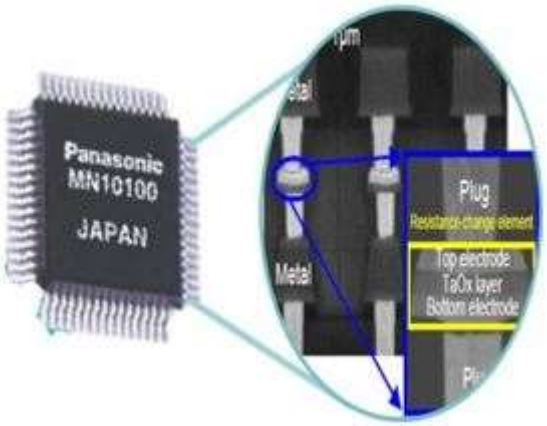


Figure I.24 : microcontrôleur commercialisé par Panasonic basé sur des OxRAM de TaO_x [81]

4.4. Contrôle du courant de compliance

Les mémoires OxRAM telles qu'elles sont schématisées en figure I.16 et I.17, c'est-à-dire constituées uniquement de la structure MIM de base, sont dites 1R (R pour résistance). En effet, aucun dispositif ne leur est associé pour contrôler le courant. Ici, la tension est directement

appliquée aux bornes de l'OxRAM. Une compliance externe est alors employée, pour empêcher un breakdown total de la cellule lors du set et du forming.

Cette structure a l'avantage d'être très simple, cependant elle ne permet pas un contrôle optimal du courant. En effet, on remarque très facilement qu'un phénomène d'overshoot de courant apparaît lors de l'étape de SET ou de forming [82, 83, 84] (cf. Fig. I.25). Ce phénomène d'overshoot est dû à des capacités parasites (notamment dû à l'utilisation de la compliance externe) et se répercute sur l'opération de reset, pendant laquelle le courant dépasse très largement la compliance utilisée durant le set. De plus le phénomène de switching est si rapide que le temps de réaction du limiteur de courant n'est pas assez court. Même si, sur les courbes en quasi-statique de set, on a l'impression que la compliance fait son travail, en réalité, le temps que la compliance externe réagisse à l'augmentation de courant, un grand courant a le temps de circuler dans le dispositif.

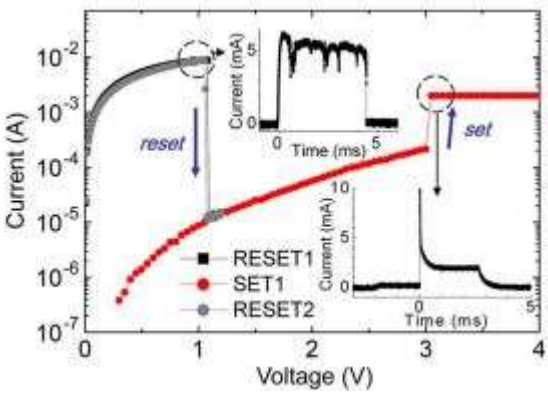


Figure I.25 : Visualisation de l'overshoot de courant lors de l'opération de SET, et de sa répercussion sur la valeur du courant de RESET. [82]

C'est pourquoi, d'autres moyens de contrôler ce courant sont nécessaires. La figure I.26 [85] compare trois moyens de contrôle de courant différents : un limiteur externe (a), intégré à l'analyseur, comme pour la figure I.25, une résistance en série avec la cellule (b) et un transistor directement implanté sur le wafer (c). Dans ce dernier cas, le courant est contrôlé via la tension de grille du transistor. On constate qu'avec celui-ci, le courant de RESET est bien inférieur, comparé aux deux premières configurations : le phénomène d'overshoot est donc largement atténué. Ceci est visible sur la figure I.27, où est représenté le courant de reset en fonction du courant de compliance utilisé durant le set. Plus les points sont proches de la courbe $I_{reset}=I_{comp}$, moins l'overshoot était important.

On a ainsi accès à un bien meilleur contrôle du courant, et des résistances de l'état hautement résistif HRS. C'est pourquoi, la structure 1T1R est actuellement considérée comme la structure référence pour les OxRAM. La plupart des travaux sur les OxRAM présentés dans la littérature s'appuie sur des structures 1T1R.

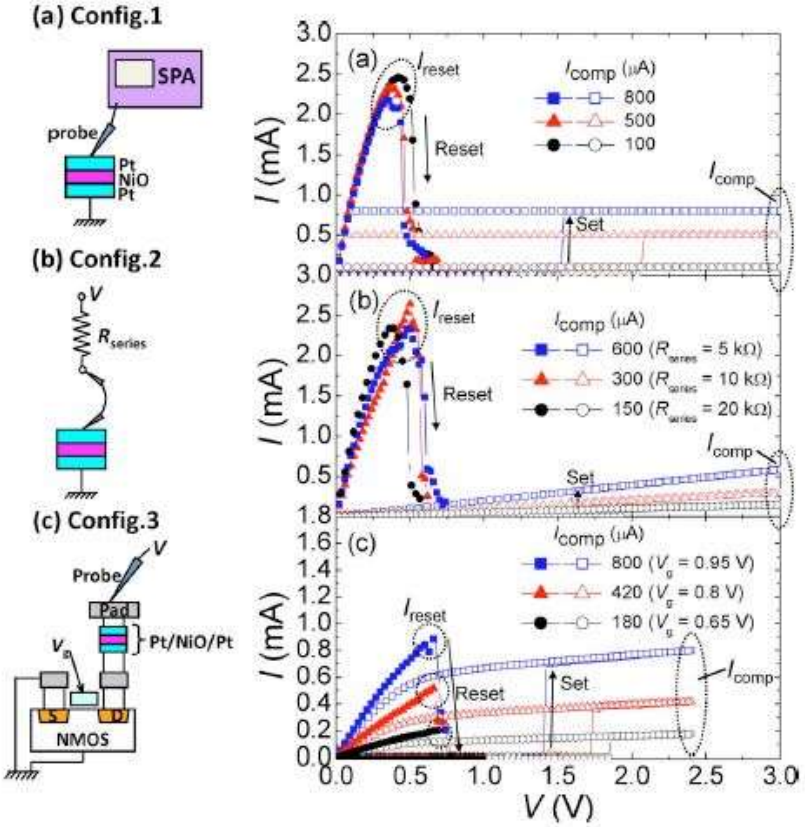


Figure I.26 : Courbes courant tension lors du reset, pour différentes configurations de test, et pour différents courants de compliance. On voit facilement que la configuration impliquant le transistor permet d'avoir des courants bien moins importants pendant le reset [85].

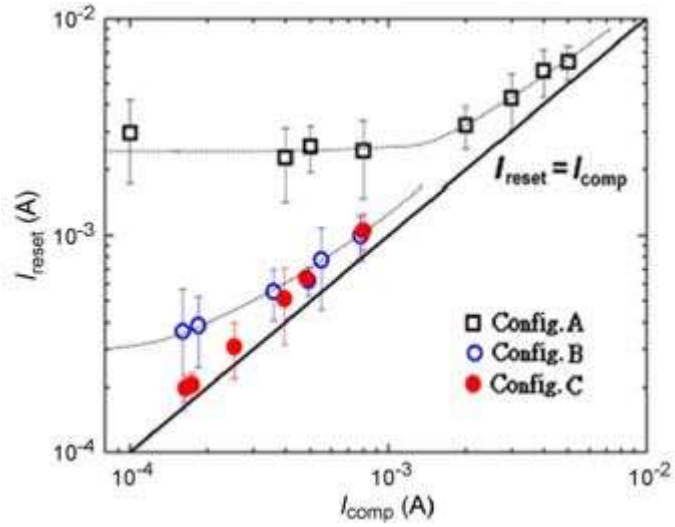


Figure I.27 : Courbe I_{reset} en fonction de I_{comp}, pour les trois configurations présentées en figure I.26 [85].

En plus de permettre de diminuer la consommation en courant, l'utilisation d'une structure 1T1R et la limitation de l'overshoot qu'elle induit permet d'améliorer considérablement l'endurance de la cellule et la valeur des résistances de l'état HRS (cf. Fig. I.28) [86].

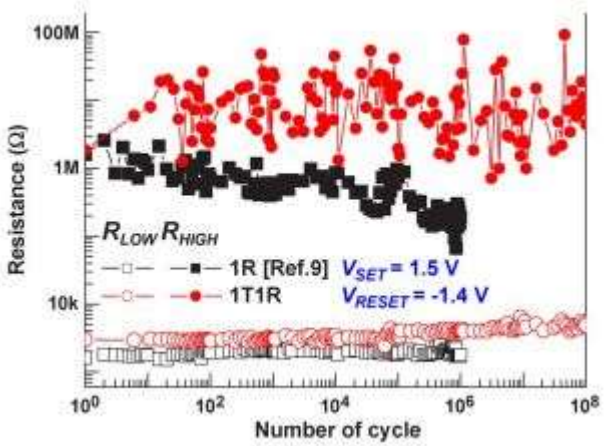


Figure I.28 : Augmentation de l'endurance et de la fenêtre résistive d'un facteur 100, via l'utilisation d'une structure 1T1R [86].

5. Etat de l'art en matière de performances

Dans cette partie, nous verrons, point par point, quelles sont les performances que l'on peut trouver dans la littérature, concernant les OxRAM. Cela nous permettra de comprendre pourquoi les OxRAM sont considérées comme parmi les candidats les plus prometteurs au remplacement des mémoires FLASH, mais également des mémoires DRAM ou SRAM.

5.1. Miniaturisation

Comme nous l'avons dit précédemment, le principal défaut des mémoires FLASH est le fait que leurs dimensions ne seront bientôt plus possible à réduire. En effet, en diminuant à l'échelle nanométrique l'épaisseur d'oxyde qui entoure la grille flottante, un courant de fuite se créé par effet tunnel et rend la cellule FLASH moins fiable. Ce problème ne se pose pas dans le cadre des OxRAM. Cependant, avec l'avènement des 3D NAND, la miniaturisation des mémoires Flash est actuellement trop avancée par rapport à celle des OxRAM. C'est pourquoi, les OxRAM visent actuellement le marché des mémoires embarquées et/ou de stockage. Il reste cependant très intéressant de savoir si la réduction en dimension de la taille des cellules a un impact sur leurs performances.

Ainsi, beaucoup d'articles s'intéressent à la possibilité des OxRAM d'atteindre des dimensions de l'ordre de 10nm, aussi bien en termes d'épaisseur que de surface.

a) Réduction de la surface

La réduction de la surface de la cellule n'a que très peu d'influence sur l'état LRS (faiblement résistif) d'une cellule [40, 87, 88, 89, 90]. Comme on peut le voir sur la figure I.29, la résistance de l'état LRS reste stable, lorsqu'on change la taille de la cellule. Cela confirme la nature filamentaire de régime de conduction en état LRS. En effet, dans la mesure où la conduction est assurée par un filament métallique au sein de l'oxyde, rien ne laisse penser que la surface de la cellule doit avoir un impact sur sa résistance. On remarque de plus que l'état HRS (hautement résistif) voit sa résistance augmenter lorsque la surface de la cellule diminue. Ceci suit approximativement la loi d'Ohm : la résistance est inversement proportionnelle à la surface. Ceci constitue un avantage pour les OxRAM. En effet, on obtient alors une meilleure frontière entre les états faiblement et fortement résistifs lorsque la surface diminue. On s'attend donc à une amélioration des performances lorsque les dimensions diminuent.

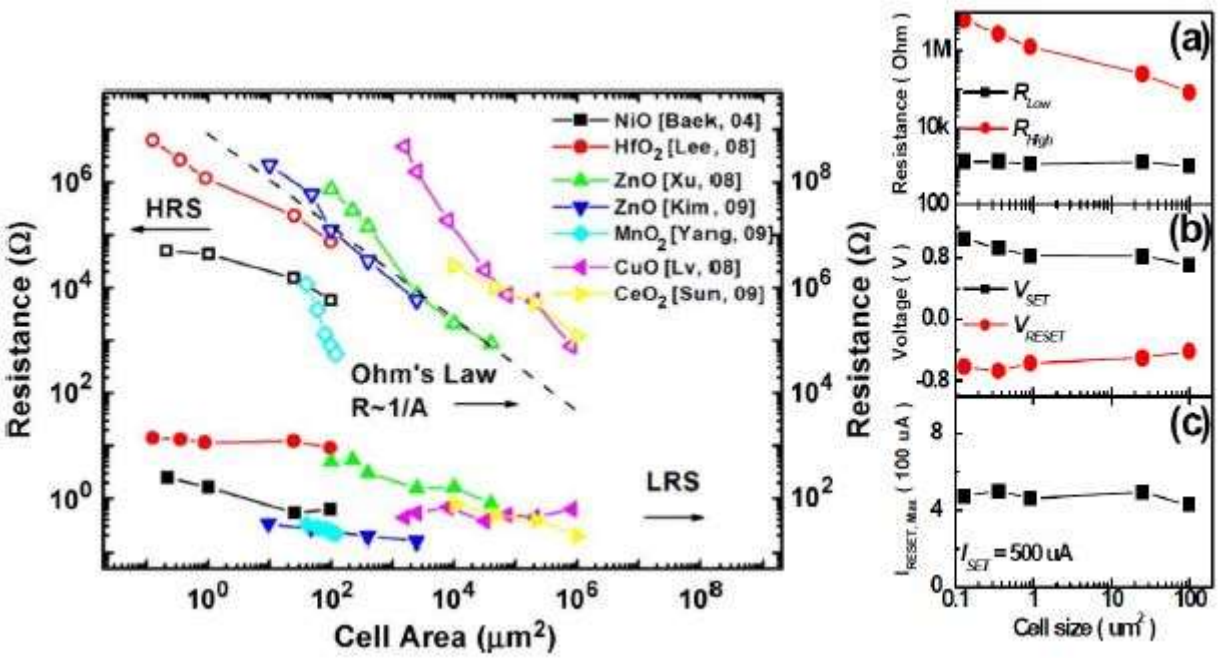


Figure I.29: Evolution des états LRS et HRS en fonction de la surface de la cellule OxRAM. L'état LRS reste globalement constant alors que l'état HRS augmente lorsque la surface diminue. [40] (gauche), [88] (droite)

Enfin, on peut noter qu'il n'y a pas d'impact de la surface de la cellule sur le courant de reset qui ne dépend que de la compliance utilisée lors du SET [40]. Actuellement, certaines cellules ont atteints des dimensions vraiment impressionnantes : dans [67], des cellules de 10x10nm² (cf. Fig. I.30) réalisent d'excellentes performances à la fois en termes de vitesse, de durabilité et de consommation : ces cellules ont des temps de commutation de l'ordre de la nanoseconde et restent opérationnelles après 5.10^7 cycles.

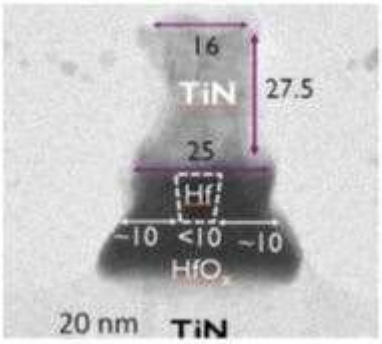


Figure I.30 : Image TEM d'une OxRAM à base d'oxyde d'hafnium, d'une largeur de moins de 10nm [67]

b) Réduction de l'épaisseur d'oxyde

La réduction de l'épaisseur d'oxyde a pour effet de diminuer la tension de forming nécessaire pour créer le filament. [88] (cf. Fig. I.31) En effet, en diminuant l'épaisseur de la cellule, si on reste à tension constante, on augmente le champ électrique. Or, c'est le champ électrique qui dirige principalement le forming des cellules. Ainsi, on peut atteindre le champ électrique nécessaire au forming à plus faible tension, si l'épaisseur est réduite. En revanche, le courant tunnel augmente, ce qui va entrainer un courant de fuite plus important. Il y a donc un trade-off entre la tension de forming et le courant de fuite.

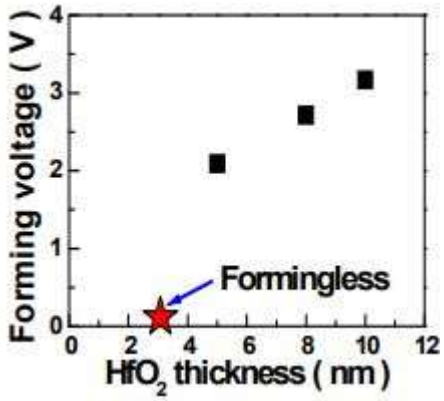


Figure I.31 : diminution de la tension de forming lorsque l'épaisseur d'oxyde diminue [88]

5.2. Endurance

L'endurance qualifie la capacité d'une mémoire à résister à un grand nombre de cycles, c'est-à-dire à un grand nombre de passage d'un état ON à OFF et inversement. Afin de quantifier cette endurance, on mesure après chaque changement d'état la résistance de la cellule. On peut alors observer l'évolution de la fenêtre résistive de la cellule. Généralement, passé un

certain nombre de cycles, certains points de mesure ne sont plus clairement définis comme correspondant à un état HRS ou LRS. Cela peut alors conduire à des erreurs de lecture. Les cellules OxRAM offrent de très bonnes performances en termes d'endurance. Certains articles parlent par exemple, avec de l'oxyde de tantale Ta, d'endurance de plus de 10^9 cycles [78]. A titre de comparaison, les mémoires FLASH atteignent entre 10^4 et 10^6 cycles, suivant les applications.

Parmi les meilleures performances en matière d'endurance on peut citer les résultats présentés par Y.B Kim et al. à la conférence VLSI de 2011 [91], où sont démontrées des endurances de plus de 10^{12} cycles obtenus sur des dispositifs à base de TaOx (cf. Figure I.32 (a)). Sur des dispositifs à base de HfO2, des cyclages de 10^{10} cycles ont été démontrés (cf. Figure I.32 (b)) [92, 93].

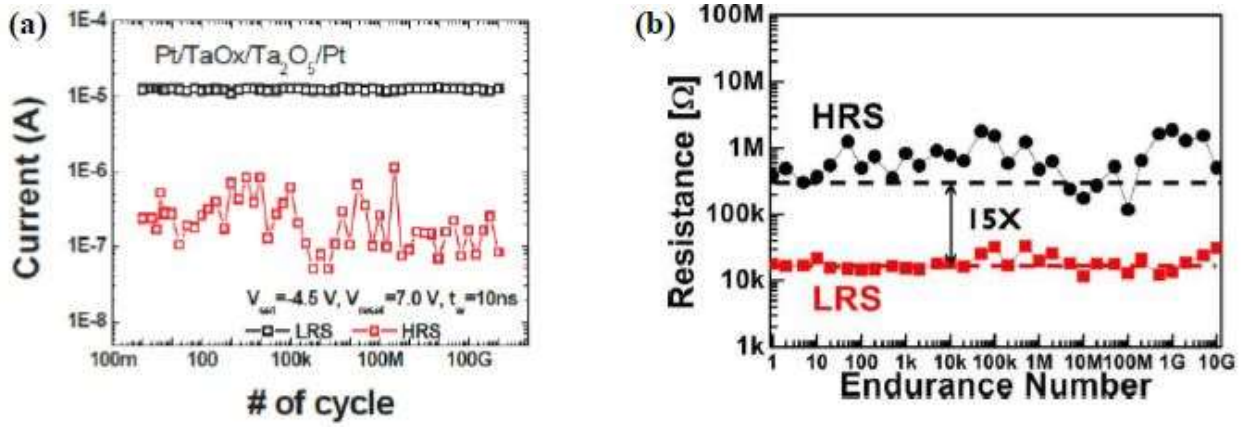


Figure I.32 : (a) Endurance de 10^{12} cycles obtenue sur des dispositifs de Ta_2O_5 [91] ; (b) Endurance de 10^{10} cycles obtenue sur des dispositifs de HfO_2 [92].

5.3. Rétention

On appelle rétention d'information la capacité d'une cellule à garder un état stable en fonction du temps. Le but est de chercher à déterminer combien de temps la cellule peut rester dans le même état. Cependant, dans la mesure où les temps recherchés sont de l'ordre de la dizaine d'année, puisqu'il s'agit de mémoires non-volatiles, il est impossible de quantifier directement la rétention d'information d'une cellule. C'est pourquoi des techniques indirectes sont mises en œuvre. Ainsi, on accélère artificiellement le vieillissement via des mesures à hautes températures. On peut ensuite extrapoler la rétention à 10 ans des cellules, via la loi d'Arrhenius (cf. Fig. I.33).

D'autres applications, notamment dans le secteur automobile, des tenues à des températures élevées, de l'ordre de 150°C, sur plusieurs sont exigées. C'est pourquoi la

rétention thermique est l'un des critères de fiabilité les plus étudiés dans la littérature, et les plus importants.

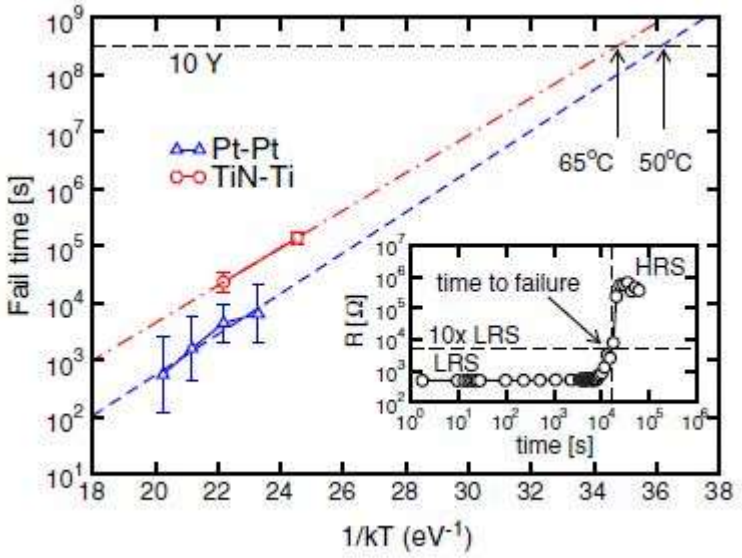


Figure I.33 : extrapolation de la température pendant laquelle l'état de la cellule est stable sur 10ans, pour des cellules avec des électrodes de Pt (50°C) et de Ti/TiN (65°C). L'encart montre la courbe qui permet de déterminer, à haute température, le temps que met la cellule avant de changer d'état (passage de LRS à HRS). [43]

De bons résultats en matière de stabilité thermique ont été démontrés sur des températures de 150°C et 200°C sur des dispositifs à base de HfO_2 [88, 94] (cf. Fig. I.34) et contribue à placer les OxRAM parmi les technologies de mémoires émergentes les plus prometteuses.

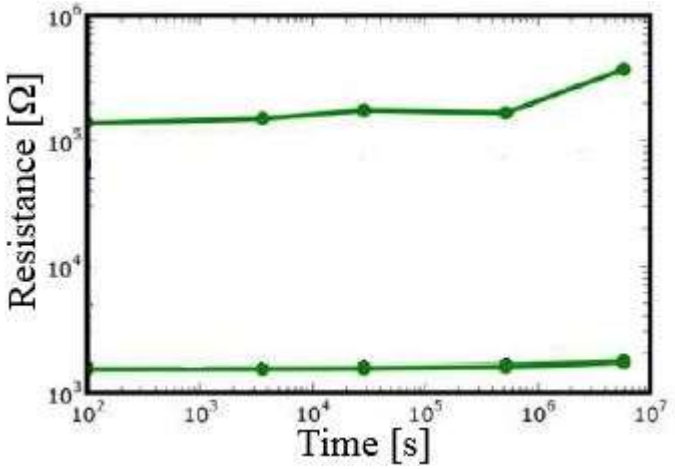


Figure I.34 : évolution de la fenêtre résistive à 150°C en fonction du temps. Au bout de 68 jours, celle-ci n'est pas dégradée, malgré une mince variation [94].

5.4. Consommation en énergie

Un autre point fort des OxRAM par rapport aux mémoires Flash est qu'elles sont opérationnelles à faible tension. En effet, les mémoires OxRAM, notamment avec électrodes

actives, nécessitent rarement des tensions supérieures à 3V. A titre de comparaison, les mémoires Flash nécessitent des tensions d'environ 17V pour les phases d'effacement.

Comme nous l'avons vu précédemment, l'ajout d'un transistor en série avec la cellule, via une structure 1T1R, permet d'avoir un excellent contrôle du courant. Ainsi, même si les OxRAM sont aujourd'hui particulièrement performantes avec des courant de l'ordre de la dizaine de µA, beaucoup de papiers font états de courant inférieurs à la dizaine de µA, voire inférieurs au µA [75, 95, 96, 97].

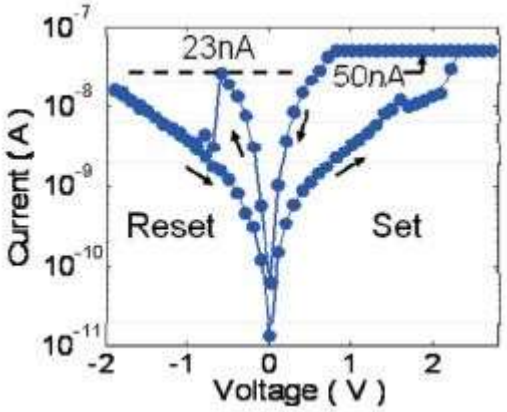


Figure I.35 : Opération de set et de reset avec des courants de 50nA, présentés à VLSI en 2011, sur des dispositifs à base d'AlO_x.

Concernant la consommation d'énergie, il faut également prendre en compte le fait que celle-ci est directement liée à la vitesse de switching de la mémoire (plus une mémoire commute vite, plus on va pouvoir utiliser des pulses courts, et donc économiser de l'énergie). La vitesse de commutation des OxRAM est traitée dans la partie suivante.

5.5. Vitesse de commutation

Dans la mesure où une grande partie du travail de thèse a été consacré à la dynamique de commutation des mémoires OxRAM, à la réalisation de mesures sur des temps ultra-courts, nous passerons plus de temps à détailler l'état de l'art concernant la vitesse d'opération des mémoires, que pour les autres types de performances.

Les mesures en quasi-statiques sont très pratiques pour vérifier le bon fonctionnement des cellules et avoir accès au courant qui circule dans le dispositif. Cependant, en pratique, pour commuter une cellule d'un état à un autre, on procède en envoyant des pulses de tension sur l'électrode désirée. Ainsi, dans le cadre d'un mécanisme bipolaire, pour réaliser un SET, on envoie un pulse sur la top électrode. Au contraire, le pulse est envoyé sur la bottom électrode lorsque l'on veut réaliser un RESET.

Réduire la taille du pulse nécessaire à la commutation de la cellule est un objectif important pour plusieurs raisons : cela permet tout d'abord d'augmenter la vitesse de fonctionnement des mémoires. De plus, cela limite l'énergie consommée. En effet, celle-ci peut s'exprimer de la façon suivante : $E_{set/reset} = \int_0^{t_{pulse}} I_{set/reset} V_{set/reset} dt$. (E étant l'énergie, I le courant, V la tension et t_{pulse} la durée du pulse).

Ainsi, dans [98], les auteurs déterminent que l'énergie de RESET diminue exponentiellement en augmentant V_{reset} (cf. Fig. I.36). L'utilisation de pulses courts semblent donc permettre de réduire l'énergie dépensée, malgré l'augmentation de la tension de reset impliquée.

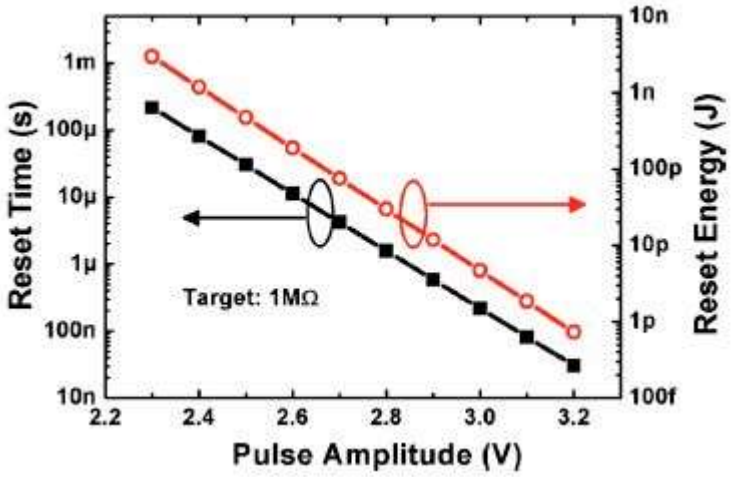


Figure I.36 : Evolution de la longueur des pulses de RESET utilisés et de l'énergie de RESET en fonction de la tension des pulses. [98]

Du point de vue des performances en matière de vitesse, les cellules OxRAM permettent un switching très rapide. Ainsi, par exemple, dans [99], un switching en 60ns est obtenu avec des cellules à base de NiO (cf. Fig. I.37).

On trouve également des publications qui parlent de temps de switching de l'ordre de la dizaine de nanosecondes [67], [88], [100] (cf. Fig. I.38). De telles performances permettraient aux OxRAM de concurrencer les mémoires DRAM, en plus des FLASH. Ces mémoires, utilisées dans la plupart des ordinateurs, permettent des temps de commutation de cet ordre de grandeur. Cependant, contrairement aux OxRAM, il ne s'agit pas de mémoires non-volatiles. Les remplacer par des mémoires répondant aux mêmes critères en matière de vitesse, mais sans la perte d'information liée à la mise hors tension, permettrait de franchir un grand cap, comme la disparition du mode veille des outils informatiques (une mémoire non-volatile n'a pas besoin d'être alimentée en permanence pour garder son contenu).

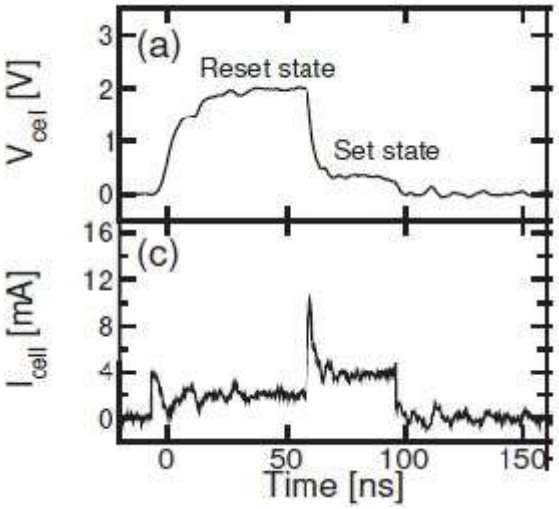


Figure I.37 : Visualisation en fonction du temps de la tension aux bornes de la cellule et du courant qui la traverse. Il y a un temps de 60ns entre l'application du pulse de tension et le changement d'état de la cellule, marqué par la brusque augmentation de courant. [99]

Cependant l'un des enjeux liés à la réduction de la taille des pulses utilisés lors des opérations de SET et de RESET, est la diminution de la fenêtre résistive lorsque le temps de pulse diminue [100]. En effet, comme on peut le voir en figure I.39, l'utilisation de pulses ultra-courts n'offre qu'une fenêtre très limitée, à moins d'augmenter fortement la tension (cf. Fig. I.40). Ce compromis à trouver entre la vitesse et la fenêtre résistive est actuellement un sujet intense de réflexion, que nous aborderons durant notre étude expérimentale.

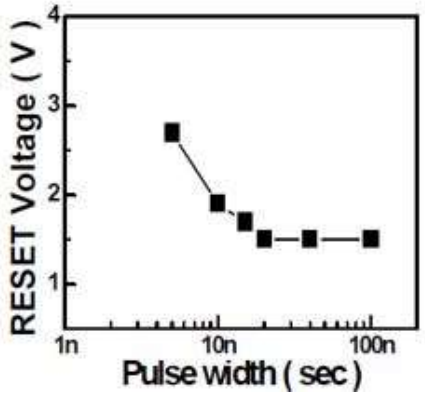


Figure I.38 : Utilisation de pulse de reset inférieurs à 10ns [88]

S'il existe un grand nombre d'autres articles, principalement issu du milieu industriel, au sujet de la réduction de temps de pulse en dessous de 10ns [88, 93, 100, 101, 102, 103], dont une performance record de 300ps [93], ceux-ci ne présentent pas d'étude de la dynamique des étapes de commutation. En effet, les faibles temps de commutation observés ne sont que des mesures de la valeur de la résistance des cellules suite aux pulses, sans observation « en direct » de l'évolution du courant. Ainsi, on ne trouve que très peu d'articles dans la littérature qui

étudient de façon dynamique les commutations sur des temps ultra-courts. A l'inverse, certains papiers, issus du domaine académique, tentent d'étudier la dynamique de switching, mais soit n'accèdent pas à des gammes de temps compétitives [104], soit n'étudient que des structures 1R, non viables en matière de maîtrise du courant [98, 99].

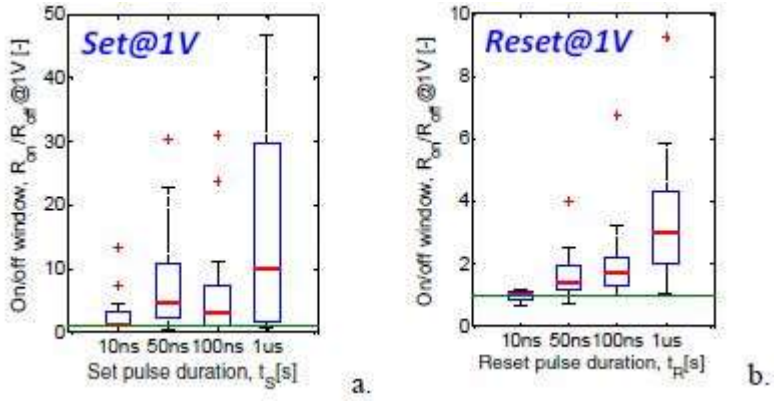


Figure I.39 : évolution de la fenêtre résistive en fonction de la durée de pulse utilisée (lors du set (a) et du reset (b)). Ce ratio Ron/Roff diminue de façon importante lorsque la taille des pulses diminue, ce qui constitue le fameux trade-off fenêtre/temps [100]

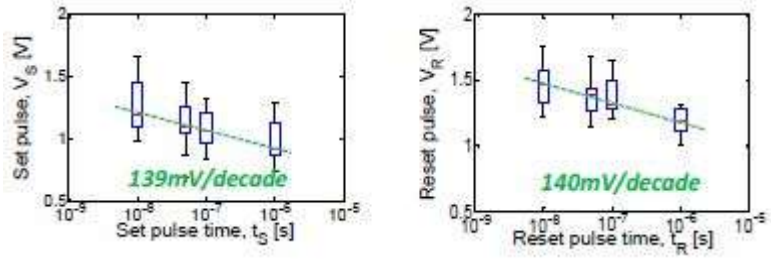


Figure I.40 : Illustration du dilemme tension/temps en et (a) et en reset (b). Pour conserver une fenêtre résistive d'un facteur de simplement 10 en diminuant la durée de pulses, il faut augmenter la tension de 140mV par décade temporelle. [100]

Ainsi l'un des principaux objectifs de ce travail de thèse sera de travailler sur des temps ultra-courts, très compétitifs, sur des structures 1T1R, adaptées aux contraintes en courant, et avec un accès à l'observation « en direct » de la dynamique de switching.

5.6. Variabilité

Une mémoire a besoin d'être capable de commuter entre deux (ou plus, dans le cadre de mémoires dites multi level) état stables, et clairement définis et distincts. Ici ces deux états sont les états LRS et HRS. L'uniformité de ces deux états est donc un paramètre crucial. Or, si les mémoires résistives démontrent des performances très compétitives dans tous les domaines listés dans les parties précédentes, elles sont particulièrement connues pour présenter une forte variabilité, aussi bien temporelle (d'un cycle à un autre) que spatiale (d'une cellule à une autre). Il s'agit là du plus gros inconvénient des OxRAM, et le principal frein à leur industrialisation.

C’est pourquoi dans cette partie nous présenterons un certain nombre d’études sur la variabilité des cellules, et essaierons de relier cette caractéristique à la physique propre des OxRAM, sans pour autant rentrer dans les détails de la modélisation du comportement des OxRAM.

La variabilité des mémoires résistives est liée au caractère stochastique du processus de switching [105-111] Comme nous l’avons dit dans la partie précédente, les étapes de formation et la destruction du filament conducteur qui permettent la commutation de la mémoire, consistent en la génération, migration et recombinaison de lacunes d’oxygène au sein de l’oxyde, qui sont des phénomènes stochastiques. Ainsi l’état LRS va dépendre du rayon de ce filament conducteur, ou du nombre de filament conducteur. La variation de l’état HRS est plus problématique. En effet, elle peut correspondre à une variation de la taille de filament rompu par l’opération de RESET. La conduction, dans l’état HRS, est souvent considérée comme dirigée par effet tunnel (du moins Trap-Assisted-Tunneling, TAT) [40, 49, 50]. Or, le courant tunnel dépend exponentiellement de la distance de filament conducteur rompu à franchir. C’est pourquoi une faible variation de la longueur de filament rompu peut conduire à une importante variation du courant tunnel développé, c’est-à-dire de la résistance de l’état HRS. Ainsi la conduction est particulièrement sensible à des variations à l’échelle atomique au sein du dispositif. La variabilité est bien une propriété intrinsèque aux mémoires résistives.

A l’International Memory Workshop (IMW) de 2013 A. Fantini présente des résultats sur l’influence de nombreux facteurs, notamment le temps de pulse, sur les variabilités cycle à cycle et device à device, sur des dispositifs TiN\HfO2\Hf\TiN (cf. Fig. I.41). D. Garbin présente également des travaux à ce sujet à EuroSOI-ULIS 2015, avec l’impact du courant de compliance sur les distributions LRS et de la tension de reset sur les distributions HRS (cf. Fig. I.42).

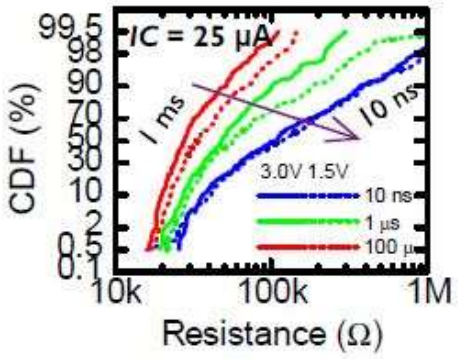


Figure I.41 : Impact du temps de pulse de set sur les distributions LRS (device-to-device sur 200 cellules) [105]

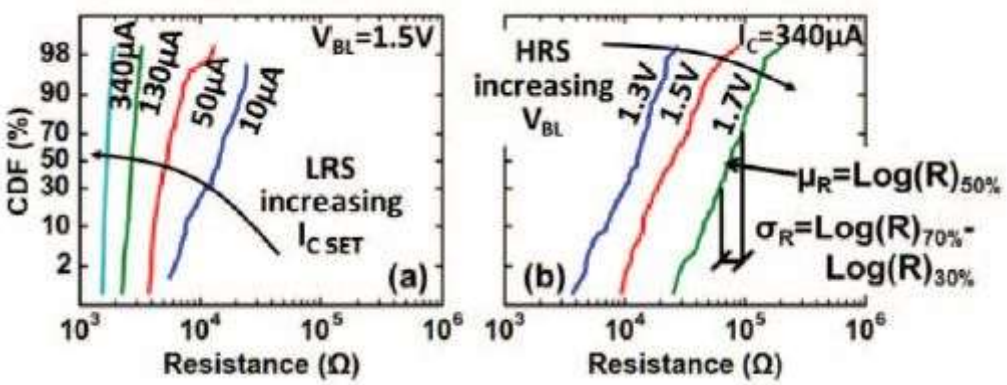


Figure I.42 : Impact du courant de compliance utilisé sur les distributions LRS (gauche) et de la tension de reset sur les distributions HRS (droite) (cycle-to-cycle sur 100 cycles set/reset) [108]

On constate facilement que la tentative de baisser le voltage, le courant ou les temps de pulse se répercutent constamment sur la qualité des distributions, spatiales et temporelles. On se heurte toujours à un compromis « performances/variabilité ».

Même si tous ces papiers présentent des explications physiques, sur les causes de ces disparités, la variabilité des OxRAM est encore un phénomène peu compris et surtout, aucune réelle solution n'existe pour pallier ce problème et permettre enfin aux OxRAM de s'affirmer en tant que candidates sérieuses et fiables pour remplacer les technologies de mémoires utilisées aujourd'hui.

5.7. Comparaison avec les autres familles de mémoires émergentes

Maintenant que nous avons situé les OxRAM en matière de performances, on peut les comparer avec les autres technologies de mémoires non volatiles émergentes.

	MAINSTREAM MEMORIES				EMERGING MEMORIES		
			FLASH				
	SRAM	DRAM	NOR	NAND	STT-MRAM	PCRAM	RRAM
Cell area	>100 F^2	6 F^2	10 F^2	<4F^2 (3D)	6~50F^2	4~30F^2	4~12F^2
Multibit	1	1	2	3	1	2	2
Voltage	<1 V	<1 V	>10 V	>10 V	<1.5 V	<3 V	<3 V
Read time	~1 ns	~10 ns	~50 ns	~10 µs	<10 ns	<10 ns	<10 ns
Write time	~1 ns	~10 ns	10 µs–1 ms	100 µs–1 ms	<10 ns	~50 ns	<10 ns
Retention	N/A	~64 ms	>10 y	>10 y	>10 y	>10 y	>10 y
Endurance	>1E16	>1E16	>1E5	>1E4	>1E15	>1E9	>1E6~1E12
Write energy (J/bit)	~fJ	~10fJ	~100pJ	~10fJ	~0.1pJ	~10pJ	~0.1 pJ

Table I.1 : Tableau récapitulatif des différentes technologies de mémoires [12]

Comme on peut le voir sur en Table I.1, les OxRAM, de même que les autres technologies émergentes surpassent les mémoires Flash dans la plupart des domaines, et peuvent également concurrencer les SRAM et DRAM sur certains critères (alors que celles-ci sont des mémoires volatiles).

Comme nous l'avons dit plus tôt, les OxRAM ne sont pour l'instant utilisées que dans des applications de niche, en tant que remplaçantes des Flash, ou des DRAM dans des contextes très spécifiques. Un travail important doit en effet être mené afin d'améliorer la fiabilité de cette technologie, et atteindre un niveau de maturité suffisant à son industrialisation.

6. Modélisation physique

Dans la mesure où une assez importante partie de ce manuscrit sera consacrée à la compréhension et à la modélisation du fonctionnement physique de commutation et de conduction des OxRAM, nous allons ici présenter quelques modèles issus de la littérature, afin de fournir les bases nécessaires à la compréhension de la physique des OxRAM. Ainsi, sans rentrer dans le détail technique de ces modèles, nous présenterons un certain nombre des différentes approches qui existent, à la fois en termes de mécanismes de commutation et de conduction.

6.1. Mécanismes de commutation

Même si le mode de commutation est sujet à de nombreux débats et questions, il est globalement accepté que la base du phénomène de commutation est basée sur la création et la destruction d'un filament conducteur. Il est également assez généralement admis que ce filament est composé de lacunes d'oxygène. Cette hypothèse est en effet confirmée par un certain nombre d'observations. En 2010, Kwon et al. ont observé au TEM, sur des mémoires à base de TiO_2, des nanofilaments conducteurs, en Ti_4O_7, qui traduisent donc une déficience en atome d'oxygène [112]. Au sein de structures en HfO_2, on peut citer le papier de Calka et al. [113], qui via une caractérisation physico-chimique assez poussée ont observé un filament conducteur d'environ 20nm. Ce filament est caractérisé, encore une fois, par une relative faible concentration en atomes d'oxygène, ce qui confirme encore que le filament est composé de lacunes d'oxygène (cf. Fig. I.43).

Cependant, l'observation de filament conducteur au sein des OxRAM est particulièrement compliquée et la stœchiométrie précise du filament conducteur n'est aujourd'hui pas encore clairement identifiée.

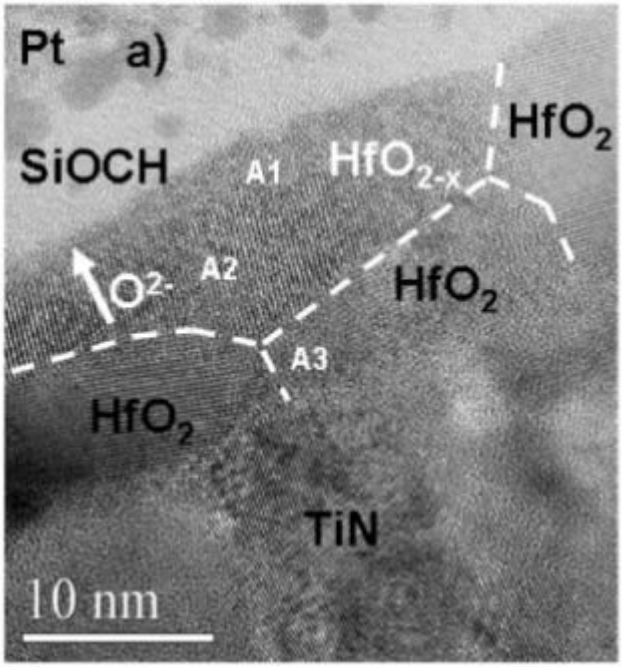


Figure I.43 : Image TEM de la zone conductrice, constituée de HfO_{2-x} [113]

Cependant, l'observation de filament conducteur au sein des OxRAM est particulièrement compliquée et la stœchiométrie précise du filament conducteur n'est aujourd'hui pas encore clairement identifiée.

On peut trier les mécanismes de commutation résistive au sein des RRAM à base d'oxyde et la formation/destruction du filament en deux catégories :

- la migration et redistribution de lacunes d'oxygène.
- la génération, migration et recombinaison de lacunes d'oxygène.

6.1.1 Mécanisme par migration de lacunes d'oxygène

Ce mécanisme, principalement représenté par le groupe de Politecnico di Milano et les travaux de D. Ielmini [37, 49, 87, 114], très souvent cités dans la littérature, suggère une redistribution des lacunes d'oxygène avec le champ électrique et la température. Dans ce modèle, la création des lacunes n'est pas décrite : les phénomènes de set et de reset sont « simplement » un déplacement de lacunes déjà existantes. Cette migration se décompose en deux composantes : une composante de diffusion, qui est dictée par le gradient de concentration et une composante de conduction, qui est dirigée par le champ électrique appliqué :

$$j_D = j_{diff} + j_{drift} = -D.\nabla n_D + \mu.F.n_D \quad (1)$$

Où j_D est le flux ($cm^{-2}s^{-1}$), D le coefficient de diffusion, μ la mobilité des ions, F le champ électrique et n_D la densité de défaut (lacunes d'oxygène).

Le coefficient de diffusion suit une loi d'Arrhenius : $D = D_0 e^{\frac{-E_A}{kT}}$ (2)

Et μ et D sont reliés par la relation d'Einstein : $\mu = \frac{qD}{kT}$ (3)

La mobilité est donc également activée selon une loi d'Arrhenius.

Pour illustrer le principe de ce modèle, prenons l'exemple d'une opération de reset. Avant l'opération, un filament conducteur relie les deux électrodes, comme représenté en Fig. I.44. Alors que la tension augmente et que le courant circule, la température augmente également par effet Joule. Ensuite, avec l'action du champ électrique les lacunes d'oxygène, chargées positivement, vont migrer vers l'électrode négative. Cette migration se produit aux points où la température est la plus élevée, puisque la migration est activée exponentiellement par la température, à la fois via μ et D. Il se créé alors une zone pauvre en lacune, qui joue le rôle de barrière isolante, du côté de l'électrode positive. Or les conductivités électriques et thermiques dépendent de la concentration en lacunes. Le champ électrique et la température augmentent alors dans cette région. La différence de potentiel se concentre essentiellement dans cette région. A mesure que la zone de déplétion augmente, l'augmentation de température et de champ électrique se fait plus localisée : en dehors de cette zone de déplétion, la température et surtout le champ électriques se font moins forts. C'est pourquoi la transition est autolimitée : la taille de la zone de rupture du filament n'augmente pas indéfiniment. L'évolution au cours du temps de la concentration en lacune, de la température et de la différence de potentiel le long de l'axe de symétrie r=0, tels qu'ils sont prédits par le modèle sont représentés en Fig. I.45.

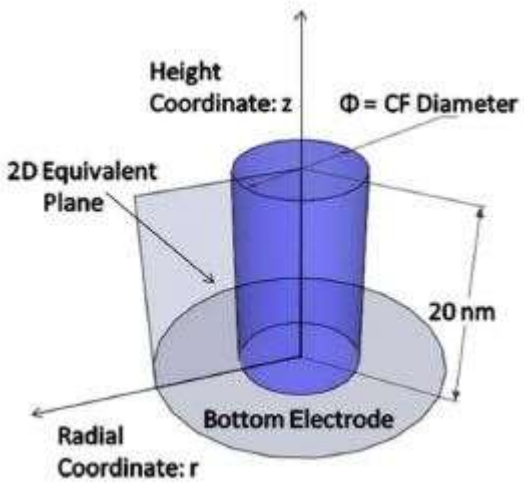


Figure I.44 : Avant le reset, un filament conducteur relie les deux électrodes. [49]

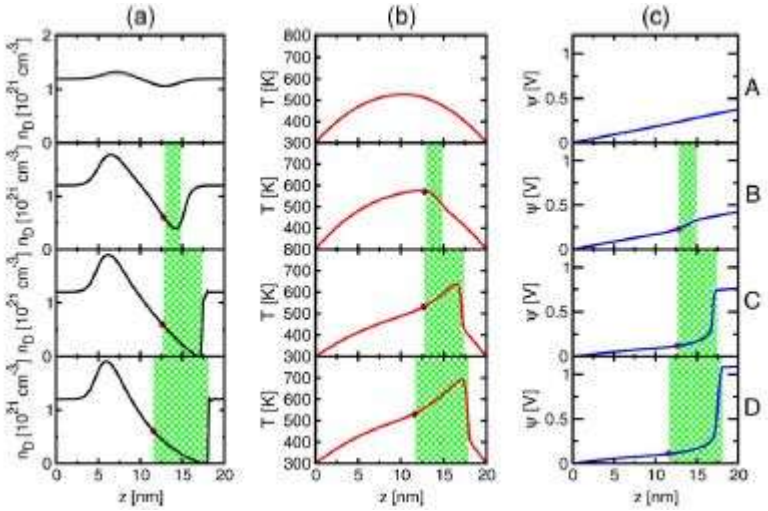


Figure I.45 : Evolution au cours du temps (de haut en bas) de la concentration en lacunes (a), de la température (b) et de la différence de potentiel (c). La zone de rupture du filament (le gap) est symbolisée en vert. [49]

Afin de mieux se représenter l'évolution du filament conducteur au cours du reset, celui-ci est représenté Fig. I.46, pour une tension de reset croissante. Sur cette image, le filament est défini comme la zone où la concentration en lacune est supérieure à $0{,}6.10^{21}cm^{-3}$. Au fur et à mesure que la tension de reset augmente, la barrière s'agrandit : la résistance augmente donc. C'est la tension de reset qui commande la taille de la zone de déplétion. Comme nous l'avons dit au-dessus, on voit bien qu'il n'y a pas de génération ni de destruction de lacunes, seulement une redistribution.

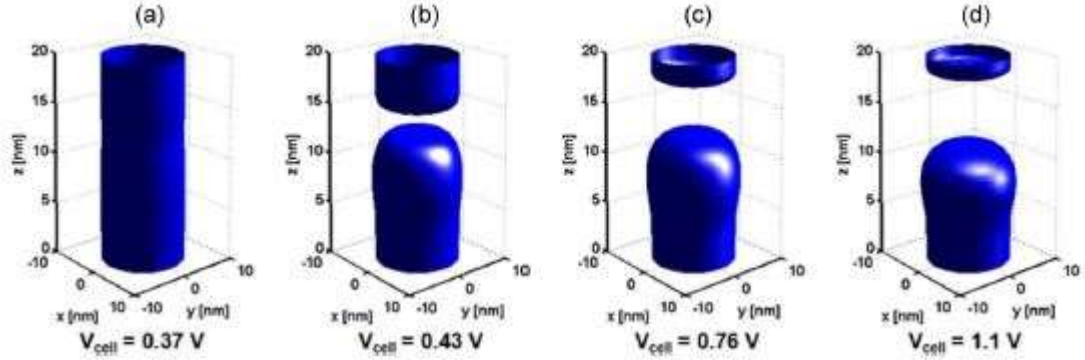


Figure I.46 : Evolution du filament conducteur au cours du temps, lorsque la tension de reset augmente progressivement. En (a) le filament est intact : on est dans l'état LRS. En (b), (c) et(d), le filament est rompu sur une distance croissante. On est dans l'état HRS. [49]

Cette approche permet de reproduire une grande partie de résultats expérimentaux, tels que des caractéristiques courant-tension, l'influence de la vitesse de rampe (de la tension de reset) sur la tension à laquelle le reset se produit, ou encore l'évolution de la résistance de reset obtenue en fonction de la tension de reset [49].

6.1.2 Mécanisme par génération, recombinaison et migration de lacunes d'oxygène

Une autre approche consiste à considérer que l'étape de set génère, via l'application d'un champ électrique des paires lacunes / ions d'oxygène. Les ions oxygène vont ensuite migrer vers l'électrode active et laisser un filament conducteur composé de lacunes d'oxygène. Pendant l'opération de reset et l'application d'un champ électrique opposé, les anions vont migrer en sens inverse et se recombiner avec des lacunes composant le filament, rompant ainsi une partie de ce dernier. Il s'agit d'une approche utilisé notamment par Stanford [115, 116], la National Chiao Tung University à Taiwan [117] ou l'Université de Pékin [118].

L'Université de Modena en Italie, et en particulier le groupe de L. Larcher a particulièrement détaillé cette approche [50, 119, 120] en se basant sur les travaux de McPherson [121-123], concernant le breakdown des diélectriques soumis à d'importants champs électriques. Ainsi, selon cette approche, la création de pairs lacunes/anions d'oxygène, opérée pendant le set, est une réaction de rupture de la liaison Hf-O, via les actions cumulées du champ électrique et de la température. Par la suite, comme avec l'approche précédente, les anions vont migrer vers l'électrode active et laisser une zone filamentaire riche en lacunes d'oxygène. Pendant le reset, les ions d'oxygène vont migrer dans le sens opposé pour ensuite se recombiner avec les lacunes et ainsi détruire une portion du filament conducteur.

Prenons l'exemple de la génération de lacune. Celle-ci est modélisée par le taux de génération suivant :

$$G(x,y,z) = \upsilon . \exp(-\frac{E_A - b.F(x,y,z)}{k.T(x,y,z)})$$

Où $\upsilon = 7.10^{13}$Hz est la fréquence de vibration effective de la liaison Hf-O [120], E_A est l'énergie d'activation de génération, à champ électrique nul, de cassure de la liaison Hf-O, F est le champ électrique local, T la température locale et $b = p_0 \cdot [(2 + k)/3]$ est le facteur de polarisation de liaison, p_0 étant le moment dipolaire moléculaire et k la constante diélectrique [121].

En figure I.47, voici un exemple de ce que cette approche est capable de prédire. Il s'agit ici de la modélisation d'un forming sur une cellule de HfO_2/Ti. Le modèle calcule la distribution des lacunes d'oxygène ainsi que des anions d'oxygène, au cours du temps. On peut ainsi constater l'influence de l'électrode active en (c) et en (i) : c'est la zone où la concentration en anions est de très loin la plus élevée, d'où son rôle de réservoir, déjà abordé dans ce manuscrit.

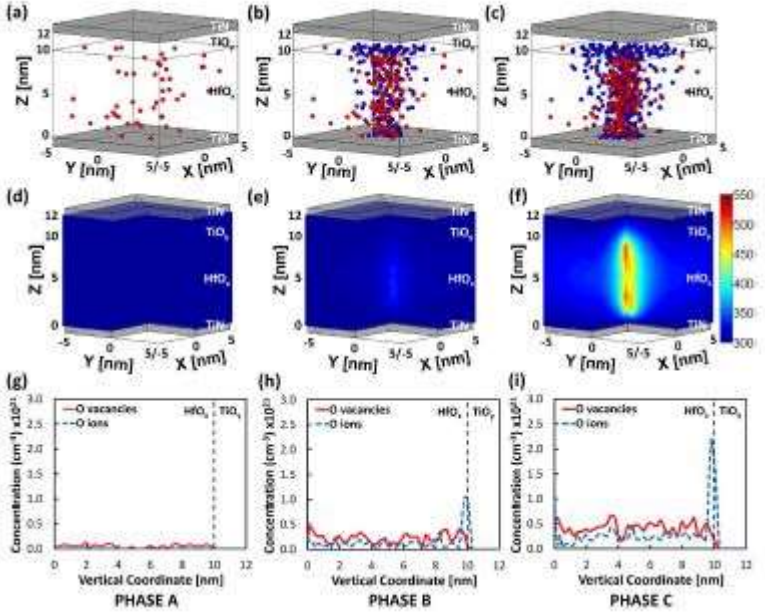


Figure I.47 : Etape de forming : (a-c) évolution de la distribution des lacunes et anions. (d-f) évolution du profil en température. (g-i) évolution de la concentration en anions/lacunes entre les deux électrodes [50]

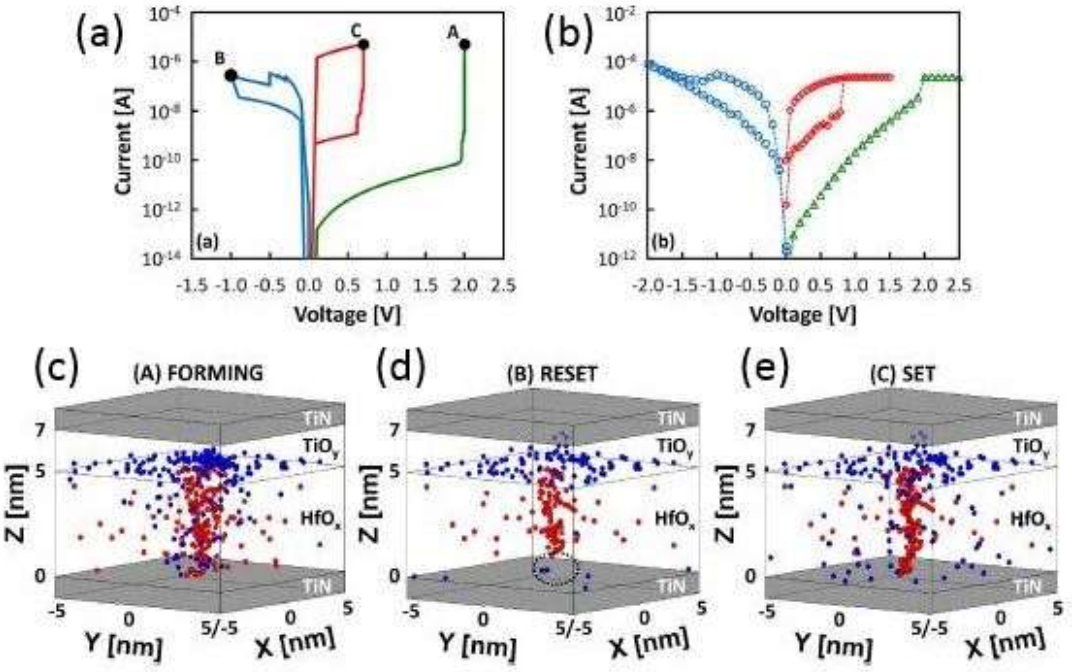


Figure I.48: (a) simulation and (b) experimental I-V curves of forming, reset and set. (c-e) distributions des lacunes/anions après les 3 opérations. [50]

Ce modèle est par ailleurs capable de modéliser des cyclages de set/reset, en partant d'une configuration post-forming (cf. Fig. I.48). Le modèle reproduit très bien les transitions abruptes (causées par la présence du terme de champ électrique en exponentielle du taux de génération) des opérations de set et de forming, ainsi que la transition plus progressive de

l'opération de reset. De plus comme on peut le voir en (d), la zone de filament rompue lors du reset se situe ici au niveau de la bottom électrode en TiN. En effet, comme le champ électrique est relativement élevé durant le reset, même si certains évènements de recombinaison se produisent durant la migration des anions vers la bottom électrode, la probabilité de recasser des liaisons Hf-O est suffisamment élevée pour que les anions atteignent la bottom électrode. A ce niveau, la diffusion des anions est bloquée par l'électrode de TiN et la répulsion par les autres ions oxygène. Ainsi, la forte concentration en ions oxygène à proximité de la bottom électrode explique que c'est à cet emplacement qu'a lieu la réoxydation du filament.

6.2. Mécanismes de conduction

S'il est globalement acquis que ce sont les lacunes d'oxygène qui sont responsable de la conduction au sein des OxRAM, le mécanisme précis n'est pas complètement cerné, d'autant plus qu'il évolue en fonction de l'état de la cellule. En effet, de façon générale, l'état LRS est décrit comme un mode de conduction métallique [49, 124]. En revanche, il y a beaucoup plus matière à débat sur l'état HRS. En effet, comme représenté sur la Figure I.49 [125], beaucoup de modes de conduction peuvent survenir.

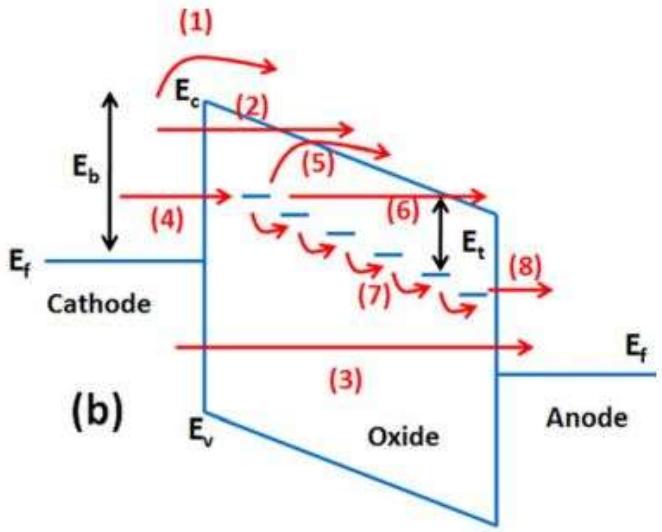


Figure I.49 : Différents modes de conduction possible : (1) émission Schottky : injection d'électron activée thermiquement vers la bande de conduction. (2) Courant tunnel Fowler-Nordheim : à fort champ, des électrons peuvent tunneler de la cathode vers la bande de conduction. (3) Effet tunnel direct de la cathode à l'anode (très peu probable). (4) Effet tunnel de la cathode à un piège. (5) Emission Poole-Frenkel d'un piège à la bande de conduction (activée thermiquement). (6) Courant Fowler-Nordheim entre un piège et la bande de conduction. (7) TAT (Trap-Assisted-Tunneling, conduction tunnel assisté par piège). (8) Effet tunnel entre un piège et l'anode [125]

Les hypothèses les plus souvent retenues sont les suivantes :

- La conduction tunnel assistée par piège [115, 118, 125]. L'Université de Modena utilise une approche plus complète de conduction assistée par pièges multi-phonons [50] (cf. Fig.I.50)

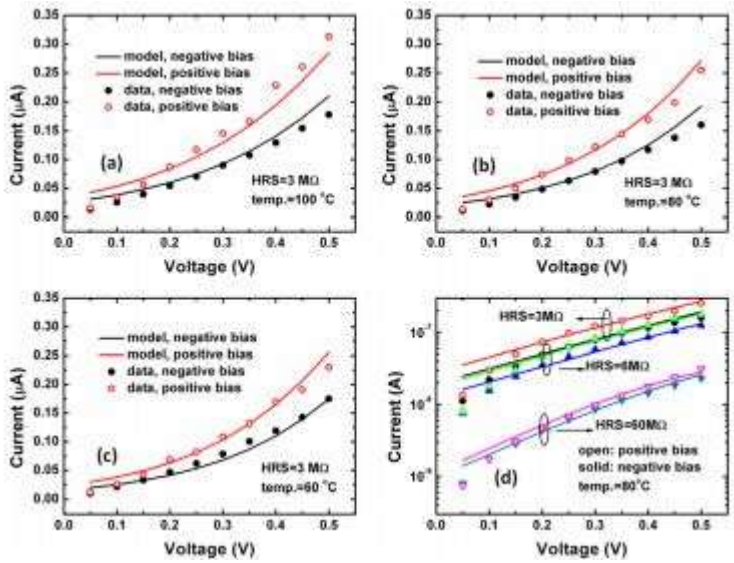


Figure I.50 : Fit de courbes I-V, via du TAT, pour différentes conditions [125]

- La conduction en Poole-Frenkel [126]. C'est notamment l'approche choisie par le groupe de Politecnico di Milano [49, 127] (cf. FigI.51)

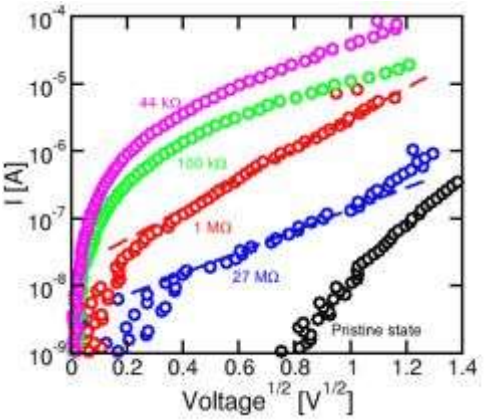


Figure I.51 : Fit de courbes I-V pour différentes résistance, de dispositifs de NiO via un courant Poole-Frenkel [127]

- Un régime de conduction par quantum point contact (QPC) est parfois utilisé, notamment par l'IMEC et permet de fitter à la fois les régimes LRS et HRS [128, 129] (cf. Fig.I.52)

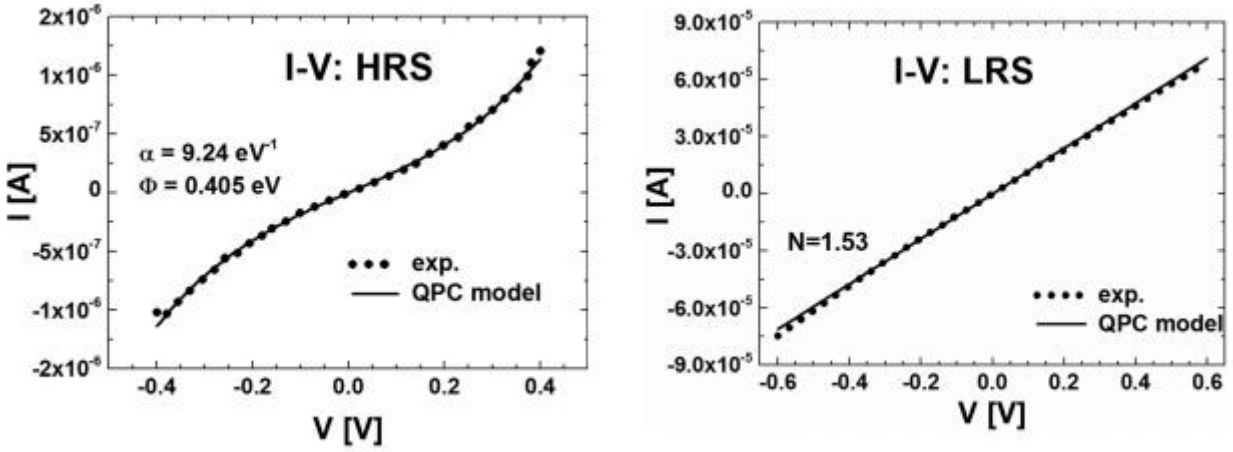


Figure I.52 : Utilisation d'un modèle de courant QPC pour fitter à la fois les régimes LRS et HRS [128]

7. Conclusion

Comme nous l'avons vu dans cette première partie du manuscrit, la technologie RRAM est particulièrement intéressante en tant que potentielle remplaçante à la technologie Flash, très largement prédominante aujourd'hui. Ses performances lui permettent également de concurrencer également les technologies de mémoires volatiles que sont les DRAM et les SRAM, sur certains aspects. De plus, leur intégration simple en Back-End-Of-Line des procédés, et leur fonctionnement à très faible tension constituent deux avantages clés par rapport aux Flash. Le mécanisme de conduction filamentaire caractéristique des OxRAM présente également l'avantage d'être économe en énergie, tout en ne limitant pas la miniaturisation de la technologie. Cependant, leur importante variabilité, ainsi que les incertitudes qui règnent sur le mécanisme réel de commutation résistives sont aujourd'hui un obstacle important à surmonter, pour atteindre une véritable démocratisation de cette technologie.

L'un de leurs meilleurs atouts est un temps de commutation très rapide, puisqu'elles sont souvent utilisées sur des temps inférieurs à 100ns. L'un des objectifs de ce travail de thèse sera d'explorer la faisabilité d'un fonctionnement sur des temps ultra-courts, de l'ordre de la dizaine de nanosecondes. Nous nous appliquerons également à étudier la dynamique du switching, sur les phases de set et de reset, sur ces temps ultra-courts, tout en étudiant l'impact de cette réduction de taille des pulses sur les distributions de résistances. Puis nous chercherons à modéliser cette dynamique, via la mise au point d'un modèle physique semi-analytique.

8. Références du chapitre I

[1] https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/

[2] http://www.martinhilbert.net/WorldInfoCapacity.html/

[3] https://www.sec.gov/Archives/edgar/data/937966/000119312514331751/d784347dex991.htm

[4] Chart courtesy of Dr. Chun g Lam, IBM Research updated version of plot from IBM Journal R&D article, 2010

[5] S. Natarajan, S. Chung, L. Paris and A. Keshavarzi, "Searching for the dream embedded memory," IEEE Solid Stat. Circ. Mag, pp. 34 - 44, 2009

[6] M.H. Kryder and C.S. Kim, "After Hard Drives—What Comes Next?," IEEE Trans. Magnetics., vol. 45, no. 10, 2009

[7] R. Bez, E. Camerlenghi, A. Modello and A. Visconti, "Introduction to Flash Memory," Proceedings IEEE , vol. 91, no. 4, 2003

[8] Q. Hubert, "Optimisation de mémoires PCRAM pour générations sub-40 nm : intégration de matériaux alternatifs et structures innovantes", 2013, Thesis dissertation

[9] J. S. Meena, S. M. Sze, U. Chand and T. -Y. Tseng, "Overview of emerging nonvolatile memory technologies," Nanosc. Resear. Lett., vol. 9, p. 526, 2014

[10] S.K. Lai, "Brief history of ETOX NOR flash memory," J. Nanosci. Nanotech., vol. 12, no. 10, pp. 7597–7603, 2012

[11] S. Gerardin and A. Paccagnella, “Present and Future Non-Volatile Memories for Space” in IEEE Transactions on nuclear science, vol. 57, no. 6, 2010

[12] S. Yu and P. Y. Chen, "Emerging Memory Technologies : Recent Trends and Prospects," in IEEE Solid-State Circuits Magazine, vol. 8, no. 2, pp. 43-56, Spring 2016

[13] https://www.extremetech.com/computing/211812-planar-nand-flash-is-dead-all-hail-our-new-3d-nand-overlords

[14] E.F. Runnion, S.M. Gladstone, R.S. Scott Jr., D.J. Dumin, L. Lie, J.C. Mitros, "Thickness dependence of stress-induced leakage currents in silicon oxide," IEEE TED, pp. 993 - 1001, 1997.

[15] International Technology Roadmap for Semiconductors (ITRS), “Emerging research devices,” 2013

[16] An Chen, “A review of emerging non-volatile memory (NVM) technologies and applications, Solid-State Electronics”, Volume 125, Pages 25-38, November 2016

[17] Y. Fujisaki, “Review of Emerging New Solid-State Non-Volatile Memories”, Japanese Journal of Applied Physics 52 (2013)

[18] https://www.eetimes.com/document.asp?doc_id=1325307

[19] T. Endoh, H. Koike, S. Ikeda, T. Hanyu and H. Ohno, "An Overview of Nonvolatile Emerging Memories— Spintronics for Working Memories," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 6, no. 2, pp. 109-119, June 2016

[20] A.V. Khvalkovskiy, D. Apalkov, S. Watts, R. Chepulskii, R.S. Beach, A Ong et al., "Basic principles of STT-MRAM cell operation in memory arrays," J. Phys. D: Appl. Phys., vol. 46, no. 074001, 2013

[21] W.J. Gallagher, "Emerging Nonvolatile Magnetic Memory Technologies," IEEE Inter. Conf. Solid Stat. Integ. Circu. (ICSICT), pp. 1073 - 1076, 2010

[22] MRAM : SERIAL AND PARALLEL MEMORY PRODUCTS, December 2013. [Online]. http://www.everspin.com

[23] B.F. Cockburn, "Tutorial on magnetic tunnel junction magnetoresistive random-access memory," Memory Tech. Desig. Test., pp. 46-51, 2004

[24] G. Prenat, G. Di Pendina, C. Layer, O. Goncalves, K. Jaber, B. Dieny et al., "Magnetic memories : From DRAM replacement to ultra low power logic chips," Design, Automation and Test in Europe Conference and Exhibition, p. 1, 2014

[25] G.W. Burr, M.J. Breitwisch, M. Franceschini, D. Garetto, K. Gopalakrishnan, B. Jackson et al., "Phase change memory technology," Journ. Vacu. Scien. tech. B., vol. 28, pp. 223 - 262, 2010

[26] D. Ielmini and A. Lacaita, "Phase change materials in non-volatile storage," Materials today, vol. 14, no. 12, 2011

[27] M. Di Ventra and Y. Pershin, "Memory effects in complex materials and nanoscale systems," Avances in Physics, no. 2, p. 60, 2011

[28] H.-S.P. Wong, S. Raoux, S. Kim, J. Liang, J.P. Reifenberg, B. Rajendran et al., "Phase Change Memory," Proceedings IEEE, vol. 98, no. 12, 2010

[29] A. L. Lacaita and A. Redaelli, "The race of phase change memories to nanoscale storage and applications," Microelec. Eng., vol. 109, pp. 351 - 356, 2013

[30] S. Yu and H.-S.P Wong, “Compact Modeling of Conducting-Bridge Random-Access Memory (CBRAM)”, IEEE Transactions on Electronic Devices, vol. 58, no. 5, may 2011

[31] C. Gopalan, Y. Ma, T. Gallo, J. Wang, E. Runnion, J.Saenz et al., "Demonstration of conductive bridging random access memory (CBRAM) in logic CMOS process," Solid-Stat. Elec., vol. 58, pp. 54 - 61, 2011

[32] Y. Gonzalez-Velo et al., “Total-Ionizing-Dose Effects on the Resistance Switching Characteristics of Chalcogenide Programmable Metallization Cells”, IEEE Transactions on Nuclear science, vol. 60, no. 6, december 2013

[33] Y. Bernard et al., “Back-end-of-line compatible Conductive Bridging RAM based on Cu and SiO2”, Microelectronic Engineering 88, 2011

[34] M.N. Kozicki, C. Gopalan, M. Balakrishnan, M. Park and M. Mitkova, “Non-Volatile Memory Based on Solid Electrolytes” in: Proceedings of the 2004 Non-Volatile Memory Technology Symposium (NVMTS), 2004

[35] K. Aratani, K. Ohba, T. Mizuguchi, S. Yasuda, T. Shiimoto, T. Tsushima et al., "A novel resistance memory with high scalability and nanosecond switching," IEEE IEDM, pp. 783 - 786, 2007

[36] E. Vianello, G. Molas, F. Longnos, P. Blaise, E. Souchier, C. Cagli et al., "Sb-doped GeS2 as performance and reliability booster in conductive bridge RAM," IEEE IEDM, pp. 31.5.1 - 31.5.4, 2012

[37] D. Ielmini, “Resistive switching memories based on metal oxides: mechanisms, reliability and scaling” Semiconductor Science and Technology, Volume 31, Number 6, 2016

[38] R. Waser, R. Dittmann, G. Staikov, and K. Szot, "Redox-based resistive switching memories – nanoionic mechanisms, prospects, and challenges," Adv. Mater., vol. 21, pp. 2632–2663, 2009

[39] G. Bersuker et al., “Metal oxide resistive memory switching mechanism based on conductive filament properties”, Journal of Applied Physics 110, 124518, 2011

[40] H.-S.P. Wong, H.-Y. Lee, S. Yu et al., "Metal–Oxide RRAM," Proceedings IEEE, pp. 1951-1970, 2012

[41] S. Yu, B. Lee, HS.P. Wong, “Metal Oxide Resistive Switching Memory”. In: J. Wu J. Cao, WQ. Han, A. Janotti, HC. Kim, “Functional Metal Oxide Nanostructures. Springer Series in Materials Science”, vol 149. Springer, New York, NY, 2012

[42] T. Cabout "Optimisation technologique et caractérisation électrique de mémoires résistives OxRRAM pour applications basse consommation", Thèse de doctorat en Micro et Nanoélectronique, 2014

[43] T. Cabout et al., “Role of Ti and Pt electrodes on resistance switching variability of HfO2-based Resistive Random Access Memory”, Thin Solid Films 533, 19–23, 2013

[44] C. Cagli, F. Nardi, and D. Ielmini, “Modeling of Set/Reset Operations in NiO-Based Resistive-Switching Memory”, IEEE Transactions on Electron Devices, vol. 56, no. 8, August 2009

[45] T.-N. Fang et al., “Erase mechanism for copper oxide resistive switching memory cells with nickel electrode,” in IEDM Tech. Dig., pp. 789–792, 2006

[46] U. Russo et al. “Conductive-filament switching analysis and self-accelerated thermal dissolution model for reset in NiO-based RRAM,” in IEDM Tech. Dig., pp. 775–778, 2007

[47] U. Russo, D. Ielmini, C. Cagli, and A. L. Lacaita, “Self-accelerated thermal dissolution model for reset programming in unipolar resistive switching memory (RRAM) devices,” IEEE Trans. Electron Devices, vol. 56, no. 2, pp. 193–200, Feb. 2009

[48] J. J. Yang et al., “Memristive switching mechanism for metal/oxide/metal nanodevices” Nat. Nanotechnol., vol. 3, no. 7, pp. 429–433, Jul. 2008

[49] Stefano Larentis et al., “Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling” IEEE Transactions on Electron Devices, vol. 59, no. 9, September 2012

[50] A. Padovani et al., “Microscopic modeling of HfOx RRAM operations: from forming to switching”, IEEE Transactions on Electron Devices, vol. 62, no. 6, June 2015

[51] B. Gao, et al., “Oxide-based RRAM switching mechanism: A new ion-transport-recombination model”, in Tech. Dig. IEEE Int. Electron Devices Meeting, pp. 563–566, 2008

[52] H. Akinaga and H. Shima, “Resistive Random Access Memory (ReRAM) Based on Metal Oxides”; Proceedings of the IEEE No. 12, December 2010

[53] K.-L. Lin, T.-H. Hou, J. Shieh, J.-J. Lin, C.-T. Chou and Y.-J. Lee, "Electrode dependence of filament formation in HfO2 resistive-switching memory," Jour. Appl. phys., vol. 109, no. 084104, 2011

[54] X. P. Wang, Y. Y. Chen, L. Pantisano, L. Goux, M. Jurczak, G. Groeseneken et al., "Effect of Anodic Interface Layers on the Unipolar Switching of HfO2-based Resistive RAM," IEEE VLSI-TSA, pp. 140 - 141, 2010

[55] C. Cagli, J. Buckley, V. Jousseaume et al., "Experimental and theoretical study of electrode effects in HfO2 based RRAM," IEEE Electron Devices Meeting IEDM, pp. 658–661, 2011

[56] A. Padovani et al., “Understanding the Role of the Ti Metal Electrode on the Forming of HfO2-based RRAMs”, Memory Workshop (IMW), 4th IEEE International, 2012

[57] J.F. Gibbons and W.E. Beadle , "Switching properties of thin NiO films," Solid-Stat. Elec., vol. 7, pp. 785-797, 1964

[58] S. Seo et al., “Reproducible resistance switching in polycrystalline NiO films,” Applied Physics Letters, vol. 85, no. 23, p. 5655, 2004

[59] I. G. Baek et al., “Highly scalable nonvolatile resistive memory using simple binary oxide driven by asymmetric unipolar voltage pulses”, in Tech. Dig. IEEE Int. Electron Devices Meeting, pp. 587–590, 2004

[60] F. Nardi, et al., “Sub-10μA reset in NiO-based resistive switching memory (RRAM) cells”, in Proc. IEEE Int. Memory Workshop, 2010

[61] L. Goux, J.G. Lisoni, X.P. Wang, M. Jurczak and D.J. Wouters, "Optimized Ni Oxidation in 80-nm Contact Holes for Integration of Forming-Free and Low-Power Ni/NiO/Ni Memory Cells," IEEE Trans. Elec. Dev., vol. 56, no. 10, 2009

[62] Z. Wei “ReRAM is Real”, Flash Memory Summit 2014 Proceedings, 2014

[63] Y.S. Chen, H.Y. Lee, P.S. Chen, C.H. Tsai, P.Y. Gu, T.Y. Wu et al., "Challenges and Opportunities for HfOx Based Resistive Random Access Memory," IEEE IEDM, pp. 31.3.1 - 31.3.4, 2011

[64] E. Vianello et al., “Resistive Memories for Ultra-Low-Power embedded computing design,” Electron Devices Meeting (IEDM), 2014 IEEE International, pp.6.3.1,6.3.4, 2014

[65] A. Benoist et al., "28nm advanced CMOS resistive RAM solution as embedded non-volatile memory," 2014 IEEE International Reliability Physics Symposium, Waikoloa, HI, 2014, pp. 2E.6.1-2E.6.5, 2014

[66] C. Walczyk, C. Wenger, D. Walczyk, M. Lukosius, I. Costina et al., "On the role of Ti adlayers for resistive switching in HfO2-based metal-insulator-metal structures_ Top versus bottom electrode integration," J. Vac. Sci. Technol. B, vol. 29, no. 01AD02, 2011

[67] B.Govoreanu et al., "10x10nm² Hf/HfOx Crossbar Resistive RAM with Excellent Performance, Reliability and Low-Energy Operation", IEEE International Electron Devices Meeting (IEDM), 2011

[68] C.H. Lien et al., "The Highly Scalable and Reliable Hafnium Oxide ReRAM and Its Future Challenges", Solid-State and Integrated Circuit Technology (ICSICT), 2010

[69] L. Zhao, Z.J. Jiang, H.-Y. Chen, J. Sohn, K. Okabe, B. Magyari-Köpe et al., "Ultrathin (~2nm) HfOx as the Fundamental Resistive Switching Element: Thickness Scaling Limit, Stack Engineering and 3D Integration," IEEE IEDM, pp. 6.6.1 - 6.6.4, 2014

[70] Y.Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini et al., "Improvement of data retention in HfO2/Hf 1T1R RRAM cell under low operating current," IEEE IEDM, pp. 10.1.1 - 10.1.4, 2013

[71] E. Vianello et al., "Back-end 3D integration of HfO2-based RRAMs for low-voltage advanced IC digital design", IC Design & Technology (ICICDT), 2013

[72] Y. S. Chen, H. Y. Lee, P. S. Chen, P. Y. Gu, C. W. Chen, W. P. Lin et al., "Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity," IEEE Electron Devices Meeting (IEDM), pp. 1 - 4, 2009

[73] S. C. Chae et al., "Multilevel unipolar resistance switching in TiO2 thin films", Appl. Phys. Lett., vol. 95, 093508, Aug. 2009

[74] C. P. Hsiunget al., "Resistance switching characteristics of TiO2 thin films prepared with reactive sputtering", Electrochem. Solid State Lett, vol. 12, pp. G31–G33, 2009

[75] Y. Wu, B. Lee, and H. P. Wong, "Al(2)O(3)-based RRAM using atomic layer deposition (ALD) with 1-mu A RESET current", IEEE Electron Device Lett., vol. 31, no. 12, pp. 1449–1451, Dec. 2010

[76] W. Kim, S. I. Park, Z. Zhang, Y.-L. Young, D. Sekar, H. P. Wong, and S. S. Wong, "Forming-free nitrogen-doped AlO X RRAM with sub-μA programming current", in Proc. IEEE Symp. Very Large Scale Integr. (VLSI), pp. 22–23, 2011

[77] Y. Hayakawa et al., "Highly reliable TaOx ReRAM with centralized filament for 28-nm embedded application" Symp. VLSI Tech. Dig. 2015

[78] Z. Wei et al., "Highly reliable TaOx ReRAM and direct evidence of redox reaction mechanism", Electron Devices Meeting, IEDM 2008

[79] T. Diokh et al., "Study of resistive random access memory based on TiN/TaOx/TiN integrated into a 65 nm advanced complementary metal oxide semiconductor technology" Thin Solid Films 533, 24–28, 2013

[80] A. Kawahara, R. Azuma, Y. Ikeda, K. Kawai, Y. Katoh, Y. Hayakawa et al., "An 8 Mb Multi-Layered Cross-Point ReRAM Macro With 443 MB/s Write Throughput," IEEE Journ. Sold-Stat. Circ., vol. 48, no. 1, 2013

[81] Panasonic Corporation [Online]. http://news.panasonic.com, 2013

[82] D.K. Kim, D.S. Suh, and J. Park, “Pulse-Programming Instabilities of Unipolar-Type NiOx”, IEEE Electron Device Letter, vol. 31, no. 6, June 2010

[83] P.-Y. Gu et al., “Scalability with silicon nitride encapsulation layer for Ti/HfOx pillar RRAM”, in Proc. Int. Symp. Very Large Scale Integr. (VLSI) Technol. Syst. Appl., pp. 146–147, 2010

[84] J. Song et al., “Effects of RESET Current Overshoot and Resistance State on Reliability of RRAM”, IEEE Electron Device Letter, vol. 35, no. 6, June 2014

[85] K. Kinoshita, et al., “Reduction in the reset current in a resistive random access memory consisting of NiOx brought about by reducing a parasitic capacitance”, Appl. Phys. Lett., vol. 93, 033506, Jul. 2008

[86] Y.S. Chen et al., “Robust high-resistance state and improved endurance of HfOX resistive memory by suppression of current overshoot”, IEEE Electron Device Letter, vol.32, no.11, November 2011

[87] F. Nardi, S. Larentis, S. Balatti, D.C. Gilmer, 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, September 2012

[88] H. Y. Lee et al., “Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO2 based RRAM”, IEEE International Electron Devices Meeting (IEDM) 2008, TECHNICAL DIGEST, 2008

[89] H. B. Lv et al., “Resistive memory switching of CuxO films for a nonvolatile memory application”, IEEE Electron Device Letter, vol. 29, no. 4, pp. 309–311, April 2008

[90] X. Sun et al., “Resistive switching in CeOx films for nonvolatile memory application”, IEEE Electron Device Letter, vol. 30, no. 4, pp. 334–336, April 2009

[91] Y.B. Kim et al., “Bi-layered RRAM with unlimited endurance and extremely uniform switching”, Very Large Scale Integr. (VLSI), 2011

[92] Y.Y. Chen, B. Govoreanu, L. Goux, R. Degraeve, A. Fantini, G.S. Kar et al., "Balancing SET/RESET Pulse for >1010 Endurance in HfO2/Hf 1T1R Bipolar RRAM," IEEE Trans. Elec. Dev., vol. 59, no. 12, 2012

[93] H.Y. Lee et al., “Evidence and solution of over-RESET problem for HfOx based resistive memory with sub-ns switching speed and high endurance”, in Tech. Dig. IEEE Int. Electron Devices Meeting (IEDM), pp. 460–463, 2010

[94] T. Cabout, L. Perniola, V. Jousseaume, H. Grampeix, J.F. Nodin, A. Toffoli et al., "Temperature impact (up to 200 °C) on performance and reliability of HfO2-based RRAMs ," IEEE IMW, pp. 116 - 119, 2013

[95] W. Kim et al., “Forming-free nitrogen-doped AlOX RRAM with sub-μA programming current”, Very Large Scale Integr. (VLSI), 2011

[96] F. Nardi et al., “Control of filament size and reduction of reset current below 10 μA in NiO resistance switching memories”, / Solid-State Electronics 58, 42–47, 2011

[97] T. Werner et al., “Spiking neural networks based on OxRAM synapses for real-time unsupervised spike sorting”, Front. Neurosci., November 2016

[98] S. Yu, Y. Wu, and H.-S.P. Wong, “Investigating the switching dynamics and multilevel capability of bipolar metal oxide resistive switching memory”; Applied Physics Letters 98, 103514, 2011

[99] D. Ielmini, C. Cagli, and F. Nardi, “Resistance transition in metal oxides induced by electronic threshold switching”, Applied Physics Letters 94, 063511, 2009

[100] Govoreanu et al., “Performance and Reliability of Ultra-Thin HfO2-Based RRAM (UTO-RRAM)”, 5TH IEEE International Memory Workshop (IMW), 2013

[101] S-S. Sheu et al., “A 5ns fast write multi-level non-volatile 1 K bits RRAM memory with advance write scheme”, Very Large Scale Integr. (VLSI), 2009

[102] K. Tsunoda et al., “Low power and high speed switching of Ti-doped NiO ReRAM under the unipolar voltage source of less than 3 V”, IEEE International Electron Devices Meeting (IEDM) 2007, TECHNICAL DIGEST, 2007

[103] L. Goux et al., " Role of the Ta scavenger electrode in the excellent switching control and reliability of a scalable low-current operated TiN\Ta2O5\Ta RRAM device”, Very Large Scale Integr. (VLSI), 2014

[104] M. G. Cao et al., “Nonlinear dependence of set time on pulse voltage caused by thermal accelerated breakdown in the Ti/HfO2/Pt resistive switching devices”, Applied Physics Letters 101, 203502, 2012

[105] A. Fantini et al., “Intrinsic switching variability in HfO_2 RRAM”, 5th International Memory Workshop (IMW), 2013

[106] R. Degraeve et al., “Causes and consequences of the stochastic aspect of filamentary RRAM”, Microelectronic Engineering 147, 171–175, 2015

[107] S. Yu, X. Guan and H.-S.P. Wong, “On the stochastic nature of resistive switching in metal oxide RRAM: physical modeling, Monte Carlo simulation, and experimental characterization”, ”, IEEE International Electron Devices Meeting (IEDM) 2011, TECHNICAL DIGEST, 2011

[108] D. Garbin et al., “Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network”, Eurosoi ULIS pp.125-128, 26-28 Jan. 2015

[109] D. Garbin et al., “HfO2-based OxRAM devices as synapses for convolutional neural networks”, IEEE Transactions on Electron Devices, vol. 62, no. 8, august 2015

[110] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy and D. Ielmini, “Statistical Fluctuations in HfOx resistive-switching memory: part I – Set/Reset variability”, IEEE Transactions on Electron Devices, vol. 31, no. 8, august 2014

[111] F.M. Puglisi and P. Pavan, “A microscopic physical description of RTN current fluctuations in HfOx RRAM”, IEEE International Reliability Physics Symposium (IRPS), 2015

[112] D.-H. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, et al., "Atomic structure of conducting nanofilaments in TiO2 resistive switching memory," Nat. Nanotechnol., vol. 5, p. 148, 2010

[113] P. Calka, E. Martinez, V. Delaye et al., "Chemical and structural properties of conducting nanofilaments in TiN/HfO2-based resistive switching structures”, Nanotech. , vol. 24, no. 085706, 2013

[114] D. Ielmini, “Modeling the universal Set/Reset characteristics of filament growth by field- and temperature- driven filament growth,” IEEE Transactions on Electron Devices, vol. 58, no. 12, pp. 4309–4317, December 2011

[115] X. Guan, S. Yu and H-S P Wong, “On the switching parameter variation of metal-oxide RRAM: I. Physical modeling and simulation methodology”, IEEE Transactions on Electron Devices, vol. 59, no. 4, April 2012

[116] S. Yu, X. Guan and H-S P Wong, “On the Switching Parameter Variation of Metal Oxide RRAM—Part II: Model Corroboration and Device Design Strategy”, IEEE Transactions on Electron Devices, vol. 59, no. 4, April 2012

[117] D. Berco and T-Y. Tseng, “A comprehensive study of bipolar operation in resistive switching memory devices”, Journal of Computational Electronics, Volume 15, Issue 2, pp 577–585, June 2016

[118] B. Gao, B. Sun, H. Zhang, L. Liu, X. Liu, R. Han, J. Kang, and B. Yu, “Unified physical model of bipolar oxide-based resistive switching memory,” IEEE Electron Device Letters, vol. 30, no. 12, pp. 1326–1328, December 2009

[119] L. Larcher, A. Padovani, O. Pirrotta, L. Vandelli, and G. Bersuker, “Microscopic understanding and modeling of HfO2 RRAM device physics,” in Proc. IEEE Int. Electron Devices Meeting, Dec. 2012

[120] A. Padovani, L. Larcher, G. Bersuker, and P. Pavan, “Charge transport and degradation in HfO2 and HfOx dielectrics,” IEEE Electron Device Letter, vol. 34, no. 5, pp. 680–682, May 2013

[121] J.W. McPherson and H. Mogul, “Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown in SiO2 thin films”, Journal of Applied Physics 84, 1513, 1998

[122] J. W. McPherson, R. B. Khamankar, and A. Shanware, “Complementary model for intrinsic time-dependent dielectric breakdown in SiO2 dielectrics”, Journal of Applied Physics 88, 5351, 2000

[123] J.W. McPherson, J. Y. Kim, A. Shanware, H. Mogul, “Thermochemical description of dielectric breakdown in high dielectric constant materials”, Appl. Phys. Lett., Vol 82(13), pp. 2121-2123, 2003

[124] M. Bocquet et al., “Robust Compact Model for Bipolar Oxide-Based Resistive Switching Memories”, IEEE Transactions on Electron Devices, vol. 61, no. 3, March 2014

[125] S. Yu, X. Guan and H-S P Wong, “Conduction mechanism of TiN/HfOx/Pt resistive switching memory: A trap-assisted-tunneling model”, Applied Physics Letters 99, 063507, 2011

[126] C. Walczyk et al., "Pulse-induced low-power resistive switching in HfO0 metal-insulator-metal diodes for nonvolatile memory applications," Journal of Applied Physics, vol. 105, no. 114103, 2009
[127] D. Ielmini, F. Nardi and C. Cagli, “Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories”, Nanotechnology 22, 254022, 2011
[128] L.M. Prócel, L. Trojman, J. Moreno, F. crupi, V. Maccaronio, R. Degraeve et al., "Experimental evidence of the quantum point contact theory in the conduction mechanism of bipolar HfO2-based random access memories," J. Appl. Phys., vol. 114, no. 074509, 2013
[129] R. Degraeve et al., “Hourglass concept for RRAM: a dynamic and statistical device model”, IEEE 21st International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA), 2014

Chapitre II : Mesures en dynamique rapides et impact sur les distributions

1. Introduction

Dans ce second chapitre, l'étude se portera sur une partie totalement expérimentale. Nous présenterons les travaux dédiés à la réduction des temps de pulse nécessaires à la commutation des cellules, ainsi qu'à l'étude de la dynamique de ces commutations.

Pour cela, dans un premier temps, nous expliquerons, de façon assez générale, les méthodes de caractérisations électriques classiques utilisées sur les OxRAM : les modes de caractérisations quasi-statiques et pulsés ainsi que les protocoles expérimentaux associés seront donc détaillés, et illustrés avec des exemples. Ensuite, nous détaillerons les deux tests de fiabilité les plus utilisés, qui sont les tests d'endurance et de rétention.

Nous présenterons ensuite les dispositifs que nous serons amenés à étudier tout au long de ce manuscrit (processus de fabrication, et particularités du réticule). En effet, ces dispositifs présentent des caractéristiques spécifiquement dédiées à l'étude en dynamique des phénomènes de commutation.

Nous avons vu dans le premier chapitre que les OxRAM sont capables de commuter sur des temps extrêmement courts, parfois inférieurs à la dizaine de nanosecondes. Après avoir confirmé la faisabilité de tels switching, avec ces dispositifs entièrement fabriqués au CEA-LETI, et avoir vérifié que ces cellules présentent des caractéristiques en matière de fiabilité du même niveau que ce qu'on peut trouver dans la littérature, nous présenterons les mesures réalisées afin d'observer de façon dynamique le switching, et ce sur une très large gamme de temps. Le Set-up expérimental dédié à ces manipulations sera également détaillé.

Enfin nous verrons quelles sont les conséquences en termes de variabilité, du fait de baisser les temps de pulses. Nous compléterons cette étude, avec une étude visant à faire varier la compliance utilisée durant le set, et la tension de reset, afin d'obtenir une analyse multiparamétrique de la variabilité des OxRAM.

2. Protocoles de mesures classiques de dispositifs OxRAM

Un grand nombre de mesures électriques ont été réalisées sur plusieurs bancs de tests. Il est donc important de consacrer un paragraphe à la description des différents types de tests réalisés sur les OxRAM. Nous allons donc décrire ici les différents modes de programmations possibles. Le premier mode de programmation est le mode quasi-statique, qui permet, sur des temps relativement longs, de s'assurer du bon fonctionnement des mémoires.

2.1. Mode quasi-statique

Les mesures quasi-statiques réalisées dans les travaux présentés dans cette partie ont été effectuées via un analyseur de semi-conducteur Keithley 4200 doté d'une unité de mesure SMU (Source Measure Unit) [1]. Le principe de la mesure quasi-statique est de réaliser des courbes I(V) aux temps longs : chaque point de mesure de courant est réalisé sur plusieurs millisecondes. Dans le cas des OxRAM, une rampe de tension est appliquée aux bornes de la cellule (1R ou 1T1R) et le courant est lu et contrôlé par l'analyseur de semi-conducteurs. Ainsi, un seul appareil permet de réaliser toutes les opérations (forming, set, reset et lecture). Même si les mesures en quasi-statiques ne permettent pas d'obtenir les caractéristiques dynamiques des mémoires, il s'agit d'un moyen simple et efficace pour déterminer les valeurs de tensions et de courant nécessaires au fonctionnement des cellules.

Dans le cadre des mesures sur des mémoires OxRAM intégrées en structure 1T1R, un port SMU est également consacré à la grille du transistor, comme représenté en Fig. II.1. Sur la grille la tension est appliquée en continue pendant la mesure, pour maintenir le transistor dans le même état. Lors de l'opération de set, une rampe de tension positive V_{top} est appliquée sur la top électrode, tandis que la bottom électrode est connectée à la masse. Comme le fonctionnement des mémoires étudiée pendant ce travail de thèse est bipolaire, le reset peut être réalisé, soit en appliquant la rampe de tension négative sur la top électrode (et donc en connectant la bottom encore à la masse), soit en appliquant une rampe de tension positive V_{bottom} sur la bottom électrode (et donc en connectant la top à la masse). Ici nous avons fait le choix de la seconde option. En effet, dans des produits industrialisables, l'emploi de tensions négatives est très contraignant : cela activerait les diodes de sécurité ESD, et nécessiterait de changer complètement l'architecture et le placement des différents transistors MOS.

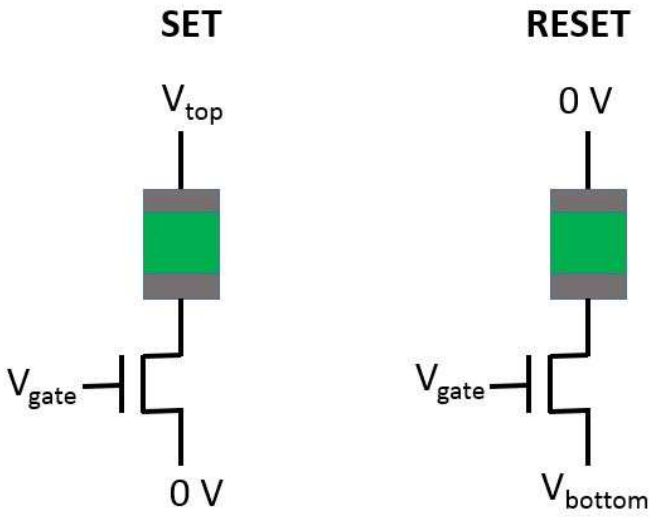


Figure II.1 : Schéma d'application des tensions pour des sets et resets sur des mémoires 1T1R. V_{gate} est une tension continue. V_{top} et V_{bottom} sont les rampes de tensions quasi-statiques.

Comme on peut le voir en Figure II.2, les mesures en quasi-statiques permettent d'appliquer la tension et de lire le courant en simultané. On a ainsi un suivi de l'état résistif de la cellule. On peut ainsi déduire des courbes quasi-statiques les tensions et courants de forming, de set et de reset. Ainsi, sur la Figure II.2, les tensions de set et reset sont respectivement de 0.55V et 0.6V. La compliance durant le set est fixée à 100µA. On peut observer que le courant de reset ne dépasse pas cette valeur, dans cet exemple.

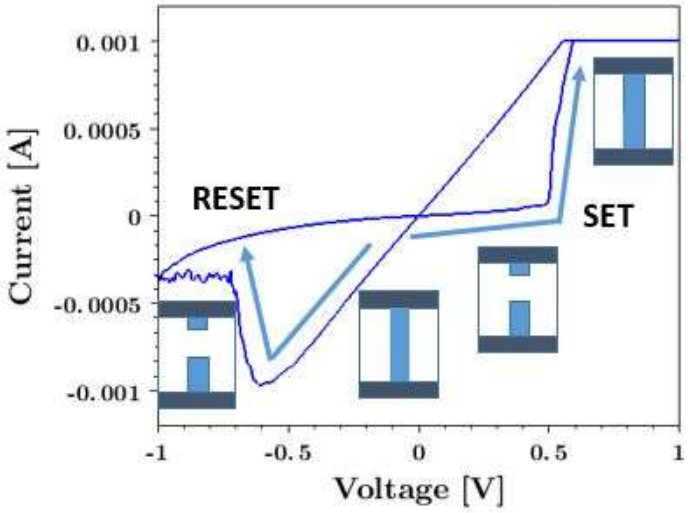


Figure II.2 : Courbe quasi-statique obtenue sur un dispositif de HfO_2/Ti. L'état supposé du filament au cours de l'opération est schématisé : le filament est reconstitué durant le set et partiellement détruit pendant le reset.

Sur cette courbe, nous avons représenté le reset par une tension négative pour des raisons de lisibilité de la courbe. En réalité le reset a été réalisé en positif sur l'électrode inférieure.

L'opération de set consiste en une brutale augmentation du courant lors de l'application d'une tension positive sur l'électrode inférieure. Cette augmentation est limitée par la

compliance. Au contraire, lorsqu'on applique une tension positive sur l'électrode inférieure, on observe à un moment donné une chute du courant.

Dans la mesure où les tests en quasi-statiques ne permettent pas de déduire les caractéristiques temporelles, et comme ce chapitre est consacré à l'étude dynamique des OxRAM, nous n'utiliserons les tests quasi-statiques que pour vérifier le bon fonctionnement des mémoires. La grande majorité de nos mesures seront effectuée en pulsé.

2.2. Mode pulsé

En mode impulsionnel, les tensions sont appliquées via un générateur de pulse. Ce type de mesure offre un accès à des données temporelles complètement inaccessibles aux mesures quasi-statiques. La forme, ainsi que la durée des impulsions peut-être contrôlée, et vérifiée via un oscilloscope.

Dans les manipulations présentées ici, nous utilisons un analyseur de semi-conducteur Keithley 4200 équipé d'un PGU (Pulse Generator Unit) 4225-RPM [2]. Les deux RPM sont connectés à la top électrode ainsi qu'à la grille du transistor. De plus, de même que pour les tests en quasi-statique, l'opération de reset est réalisée via une tension (en pulsé ici) positive sur la bottom électrode. Nous utilisons donc un autre générateur de pulse HP 8110A [3] connecté à la bottom électrode. Ainsi, nous n'avons pas besoin de changer les connections entre les phases de set et de reset, ce qui permet de réaliser des mesures de cyclages. De plus, en appliquant une tension négative sur l'électrode supérieure, on risque de polariser en direct les diodes PN drain-bulk et source-bulk, et donc de court-circuiter le transistor, dont le courant ne sera plus contrôlé par la grille.

La taille des impulsions peut être réglée sur une importante plage d'ordres de grandeur, allant de plusieurs millisecondes à quelques nanosecondes. Après chaque pulse de set et de reset, une opération de lecture est effectuée, afin de lire la résistance de la cellule après l'opération. L'opération de read consiste à appliquer une faible tension (en général de 100mV, afin de ne pas modifier l'état résistif de la cellule) sur la top électrode, pendant un certain temps (par défaut les temps de lectures sont de l'ordre de la dizaine de microseconde). L'analyseur mesure le courant plusieurs fois durant la durée de ce pulse et calcule la moyenne des points. Le 4225-RPM que nous utilisons permet une résolution en courant inférieure à 200pA et une résolution temporelle de 20ns. Une représentation d'un train de pulse classique, lors d'un cycle set/read/reset/read est schématisée en Figure II.3.

Comme les structures étudiées sont des structures 1T1R, durant chaque opération, en parallèle aux pulses de set et reset appliqués aux top et bottom électrodes, un pulse est appliqué

sur la grille du transistor, afin d'autoriser le passage du courant. Durant le set, cette tension permet de contrôler le courant de compliance. Durant le reset et les lectures, une tension plus importante est appliquée, afin d'ouvrir au maximum le transistor, et d'abaisser au maximum sa résistance.

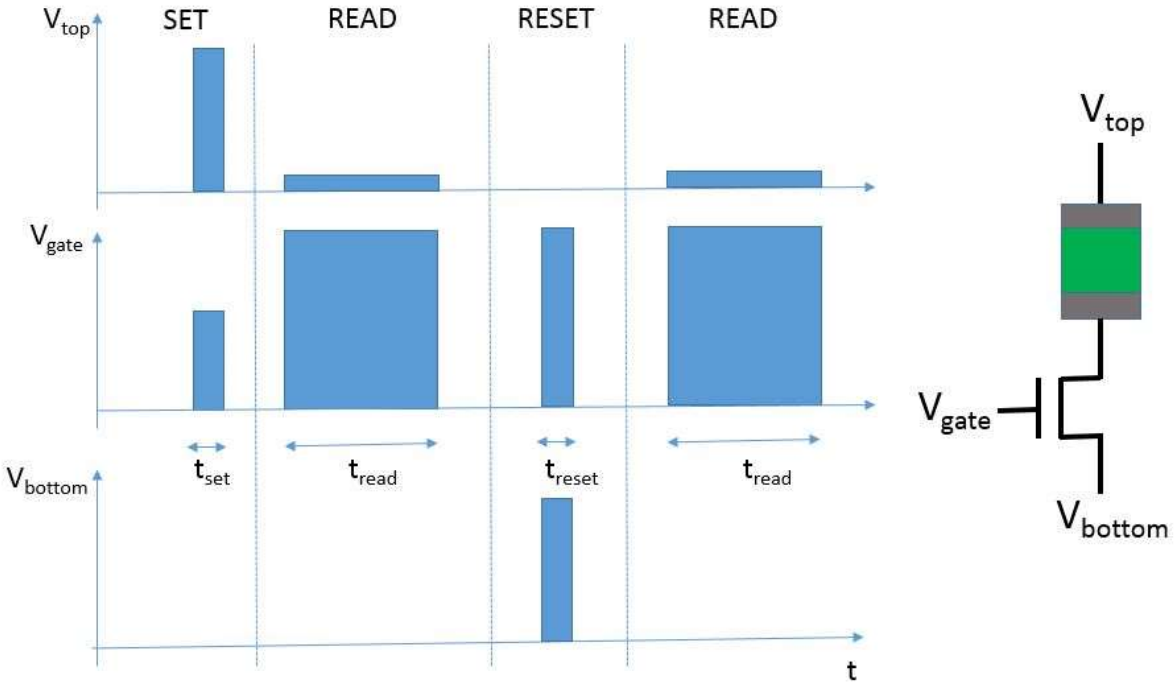


Figure II.3 : Séquence de pulses rectangulaires utilisée pour la programmation de cellules OxRAM.

Les données de sorties des mesures en pulsé sont principalement les valeurs de résistances des états LRS et HRS en fonction du numéro de cycle correspondant. En Figure II.4, un exemple de lecture de résistance, pour dix cycles de set/reset. On peut ainsi suivre l'évolution des états résistifs cycle après cycle.

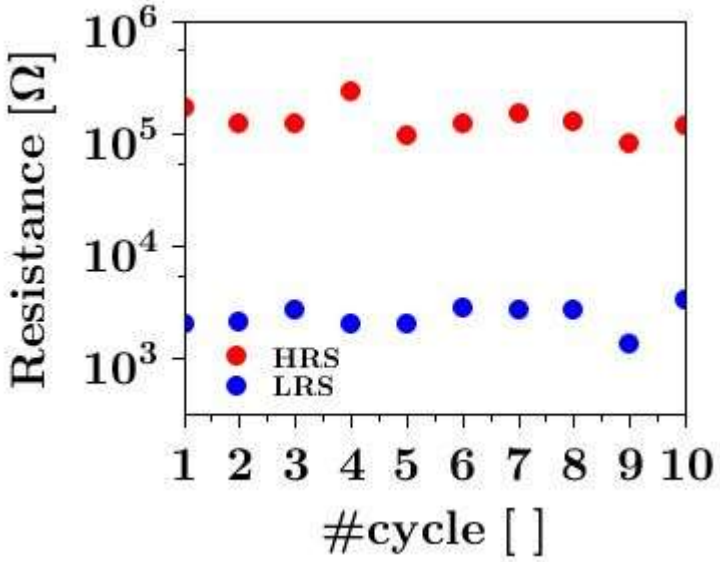


Figure II.4 : Exemple d'une série de 10 cycles en pulsé effectuée sur une mémoire de HfO_2/Ti. Une lecture de résistance est effectuée après chaque opération.

2.3. Mesures d'endurance

Réaliser un cycle complet de set/read/reset/read en pulsé est une opération particulièrement rapide (bien évidement dépendant des paramètres utilisés). Il est donc aisé de réaliser des mesures d'endurance, c'est-à-dire un nombre plus importants de cycles d'écriture-effacement. Les tests d'endurance permettent de déterminer le nombre de cycles d'écriture-effacement que l'on peut faire subir à une cellule avant soit qu'elle ne cesse de fonctionner, soit que la fenêtre résistive devienne trop faible.

Dans la mesure où on peut être amené à réaliser un très grand nombre de cycles (régulièrement 10^7 et 10^8 cycles), les opérations de read ne sont pas réalisées après chaque cycle, à la fois pour économiser du temps, et pour ne pas avoir à traiter des gigabits de données pour une seule cellule. Ainsi, celles-ci sont réalisées de façon logarithmique, tout au long de la mesure d'endurance, comme c'est le cas sur la mesure représentée en Figure II.5. Ce test a été réalisé avec des temps de pulse de set et reset de 10µs, des temps de lecture de 20µs et des temps d'attente entre les opérations de 1µs. Ainsi, la durée totale du test est de 25s. La gamme de courant utilisé est le range 100µA, le sample rate est de 200 millions de points par seconde.

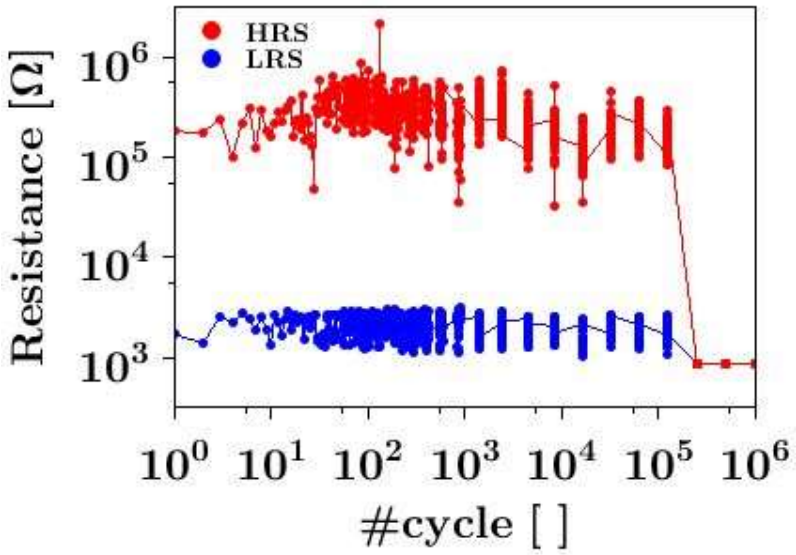


Figure II.5 : Exemple de mesure d'endurance réalisée sur des dispositifs de HfO_2/Ti. Chaque point correspond à une mesure de résistance à un numéro de cycle donné. C'est pourquoi on obtient cette allure en nuage de points. Comme très souvent, l'état HRS est plus dispersé que l'état LRS. On observe ici, qu'après un peu plus de 100 000 cycles, la cellule devient incapable de s'effacer et reste bloquée dans l'état LRS. La cellule est dite cassée.

L'International Technology Roadmap for Semiconductors (ITRS) impose un objectif de 10^7 cycles pour les technologies de mémoires émergentes, pour envisager un remplacement des technologies Flash [4].

2.4. Distributions résistives et variabilité

En général, afin de mieux se représenter la distribution des résistances, on trace plutôt les CDF (Cumulative Distribution Function). Ces distributions peuvent être des distributions temporelles (c'est-à-dire, sur une même cellule, réaliser un grand nombre d'écriture/lecture/effacement/lecture, et analyser la variabilité cycle-à-cycle. La démarche est ainsi semblable à des mesures d'endurance. En général, les deux traitements de données sont réalisés en parallèle), ou des distributions spatiales (c'est-à-dire ne réaliser que quelques cycles, mais sur un grand nombre de cellules, et ainsi analyser la variabilité cellule-à-cellule). Le second type de mesure nécessite d'avoir à disposition des matrices de mémoires (au minimum de quelques kilobits, afin d'avoir un échantillon statistique suffisant) et un banc de test adapté, et sont plus représentatives d'une matrice de dispositifs telle qu'on pourrait l'intégrer dans un produit . Un exemple de CDF est représenté en Figure II.6. Ce type de tracé est intéressant pour visualiser globalement des distributions. Néanmoins, il ne permet pas de bien traiter les « queues de distribution », c'est-à-dire les points les plus en retraits de la valeur médiane. Or ce sont ces points qu'il convient de mettre en valeur lorsque l'on veut s'assurer de la fiabilité de dispositifs.

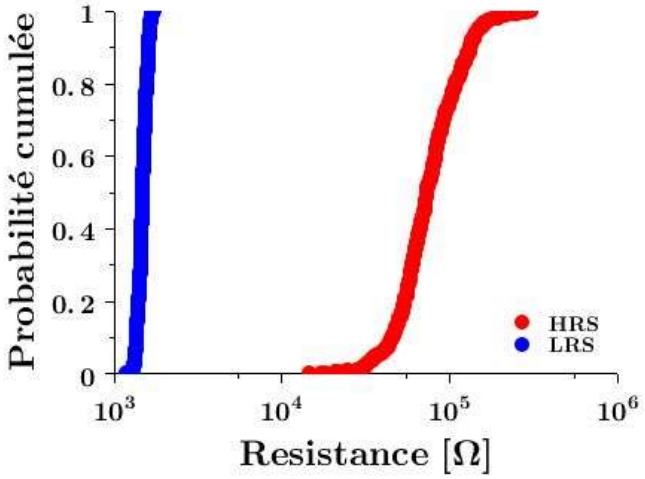


Figure II.6 : Représentation de distribution cumulée de 1500 lectures, réalisées de façon logarithmique durant un cyclage de 10^6 cycles sur un dispositif de HfO2/Ti.

C'est pourquoi, il est d'usage de représenter ces distributions à l'échelle de déviation standard (droite de Henry) : pour chaque probabilité cumulée, on peut associer un multiple de la déviation standard σ calculé à partir de la loi log-normale. Cette représentation permet d'obtenir une vision logarithmique de la distribution et de chiffrer plus facilement le taux d'erreur (cf. Fig.II.7). En fonction de la fenêtre résistive désirée (c'est-à-dire le ratio R_{HRS}/R_{LRS}), on peut déterminer la taille maximale de la matrice que l'on peut construire. Par

exemple, un croisement des courbes à 5σ (soit environ 1 point sur 1 million) est nécessaire pour s'assurer des matrices Mbits fiables.

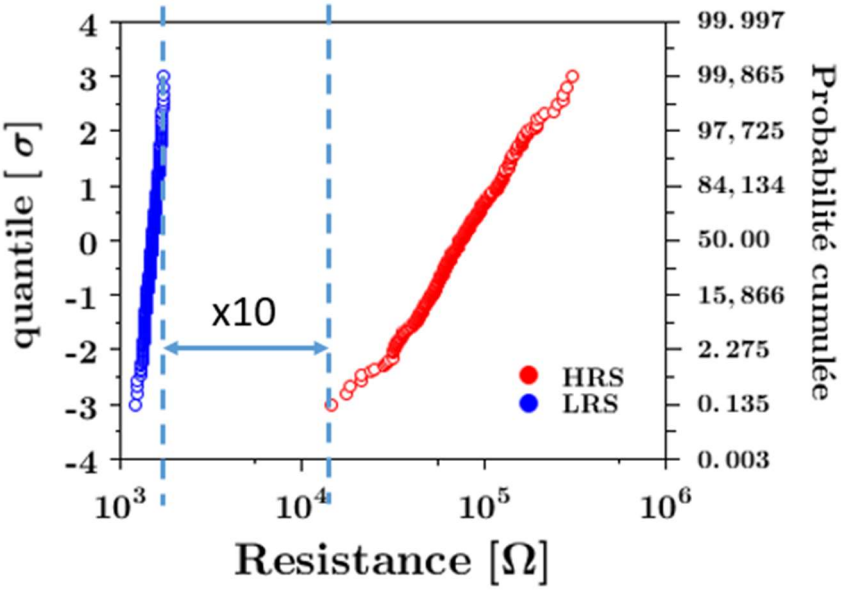


Figure II.7 : Représentation des mêmes distributions mais à l'échelle de déviation standard. On remarque qu'on conserve un facteur 10 entre la lecture LRS la plus élevée et la lecture HRS la plus faible, à 3σ (soit les 0.3% les plus éloignés de la valeur médiane). Comme pour la figure précédente, on constate que l'état HRS est plus dispersé que l'état LRS. Néanmoins, dans ce cas-ci, les deux distributions sont à peu près normale (assimilables à des droites). En axe vertical secondaire, nous avons représenté les probabilités cumulées correspondante à chaque multiple entier de σ.

3. Présentation du réticule MARS (Mémoire Avancée Résistive à Sélecteur) et des différents splits

Afin de réaliser des mesures dynamiques, sur des temps ultra-courts, nous avons utilisé un véhicule de test OxRAM réalisé intégralement dans les salles blanches du CEA-LETI. Ce véhicule de test comprend à la fois des structures 1R et des structures 1T1R, avec différentes dimensions de transistors. De plus, différents splits ont été réalisés, avec plusieurs matériaux utilisés en tant qu'électrodes supérieures. L'oxyde reste quant à lui de l'oxyde d'hafnium. De plus, comme nous le verrons dans la description du réticule, certains éléments mémoires présentent des particularités permettant l'optimisation de la mesure des paramètres dynamiques que l'on cherche à obtenir.

3.1. Présentation du réticule

Ce réticule, prénommé MARS, est destiné à l'intégration de dispositifs OxRAM à base d'oxyde de hafnium, mais également PCRAM ainsi que CbRAM, en BEOL (Back-End-Of-Line). Les empilements mémoires sont intégrés entre les niveaux de métal 1 et de métal 2(cf. Fig. II.8 (a)). Dans le cas des structures 1T1R, l'électrode inférieure (juste au-dessus du Métal 1) est directement implantée sur le drain du transistor, via un pilier en tungstène (en rouge sur la figure II.8 (b)). Les lignes de Métal 1 et 2 se croisent selon une structure dite Cross-bar, et l'empilement mémoire est située à l'intersection entre ces deux lignes (cf. Fig. II.8 (c)). Enfin, les lignes de métal sont connectées à des plots de surface, permettant les mesures sous pointes, pour les caractérisations électriques.

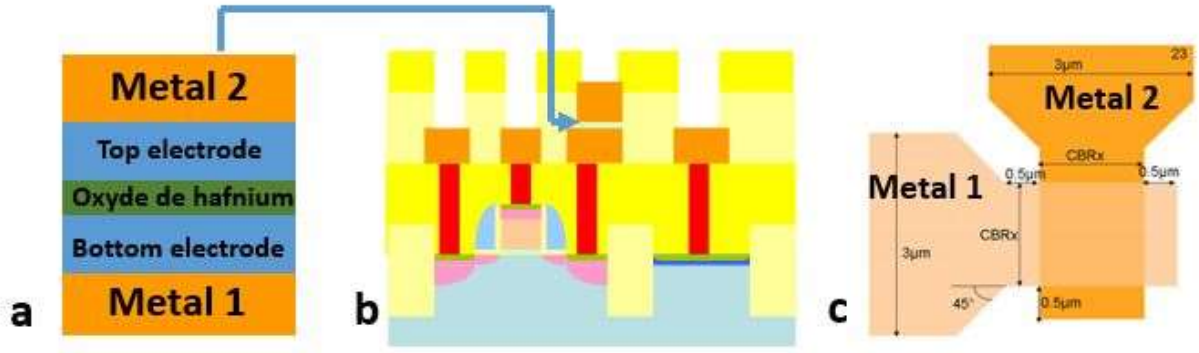


Figure II.8 : (a) Schéma de l'empilement mémoire. (b) Schéma de l'intégration de la strucutre MIM au sein du dispositif complet. (c) Représentation schématique de la structure en cross-bar des deux lignes de métal 1 et 2.

Plusieurs splits ont été fabriqués, avec différentes épaisseurs des oxydes et électrodes, ainsi que différentes surfaces de la mémoire ont été testées. Celles-ci seront précisées lorsque nous aborderons les résultats expérimentaux.

Enfin, ce réticule présente une structure particulière, à 5 plots (en temps normal, seuls 4 plots sont nécessaires, comme indiqué en Fig. II.8 (b) : top electrode, grille du transistor, source du transistor et prise de masse). Dans la structure à 5 plots, un cinquième plot est intégré entre le drain du transistor et l'empilement mémoire, comme schématisé en Figure II.9. Le but de ce cinquième plot est de mesurer la tension exacte aux bornes de la mémoire (à la place de la tension totale aux bornes du circuit série MOS-OxRAM). Comme nous le verrons, c'est ce cinquième plot qui nous permettra de réaliser nos mesures sur la dynamique de switching des mémoires.

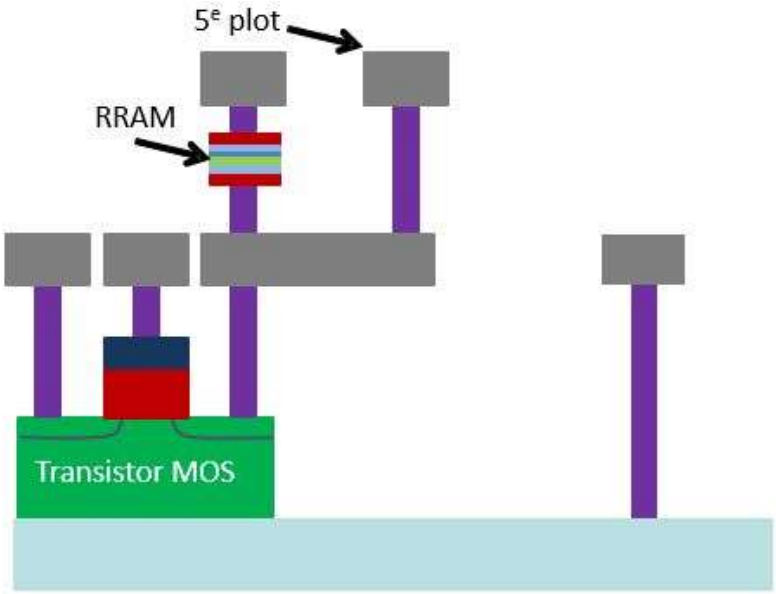


Figure II.9 : Vue schématique décrivant la position du plot permettant de sonder la tension aux bornes du dispositif OxRAM uniquement.

3.2. Fabrication des dispositifs

Les dispositifs MARS sont tous fabriqués intégralement au LETI. Dans cette partie nous détaillerons le flot d'intégration de l'empilement mémoire, à partir de la couche de Métal 1.

(a) Le niveau de Métal 1 est déposé par PVD (Physical Vapor Deposition). Il est composé d'un empilement de 10 nm de Ti, 440nm d'AlCu, 10 nm de Ti et de 100 nm de TiN. Le rôle des deux couches de Ti et TiN est de permettre une meilleure adhésion à la fois avec le pilier en tungstène et l'empilement mémoire. La couche en nitrure de titane TiN constitue l'électrode inférieure de l'empilement mémoire. Cette déposition par PVD est suivie d'une lithographie, d'une gravure ainsi qu'un dépôt de passivation en oxyde de silicium SiO_2. Pour finir cette première étape, une planarisation de la surface est réalisée par CMP (Chemical Mechanical Planarization).

(b) Ensuite, la couche d'oxyde d'hafnium (5nm ou 10 nm) est déposée par ALD (Atomic Layer Deposition). L'ALD est une technique de déposition de couches minces. Le principe consiste à exposer une surface successivement à différents précurseurs chimiques afin d'obtenir des couches ultra-minces, de quelques nanomètres d'épaisseur. Dans ce cas, le dépôt est réalisé à 300°C. Les précurseurs, pour déposer du HfO_2 sont le tétrachlorure d'hafnium $HfCl_4$ et de l'eau. C'est le nombre de répétition d'exposition à ces précurseurs qui fixe l'épaisseur de couche obtenue.

Par la suite, l'électrode supérieure (c'est-à-dire l'électrode active) est déposée par PVD au-dessus de l'oxyde. L'épaisseur de cette couche est de 10nm, quel que soit le matériau utilisé. Au-dessus, toujours par PVD, une couche de TiN de 50nm est déposée.

(c) Le niveau de Métal 2 Ti/AlCu/Ti/TiN, identique au niveau Métal 1 est déposé par PVD.

(d) Pour finir, une lithographie et une gravure de l'oxyde, de l'électrode supérieure et de la couche de Métal 2 sont réalisées.

Toutes ces étapes sont résumées en Figure II.10.

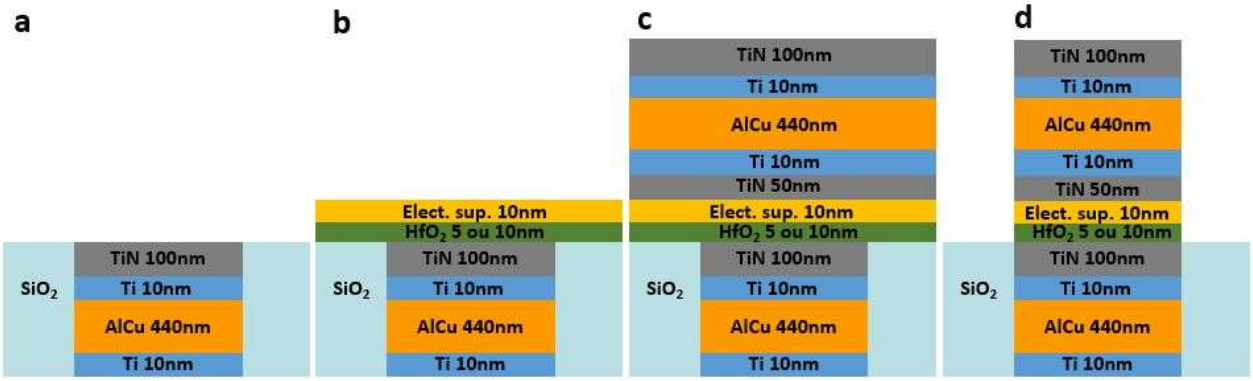


Figure II. 10 : Schématisations des différentes étapes technologiques de la fabrication des dispositifs mémoires.

4. Etude dynamique du switching sur des temps ultra-courts

L'objectif de cette partie est de présenter les mesures dynamiques réalisées. Pour ce faire, nous présenterons tout d'abord le banc de test utilisé. Ensuite nous passerons à la caractérisation du véhicule de test présenté dans la partie précédente : après avoir démontré que ce véhicule de test est bien fonctionnel, nous passerons au cœur de l'analyse dynamique, avec l'utilisation du cinquième plot présenté précédemment, permettant de déterminer avec une grande précision le moment exact de la commutation.

4.1. Présentation du set-up

Dans cette partie nous détaillerons les caractéristiques du banc de test électrique dédié à ce véhicule de test OxRAM.

Le banc de test utilisé est un banc complètement dédié à la caractérisation électrique sous pointes de mémoires non-volatiles. Ce banc sera utilisé à la fois pour les mesures en quasi-statique et les mesures impulsionnelles.

Le banc de test est équipé d'un analyseur de semiconducteurs Keithley-SCS, doté à la fois d'une unité de mesure SMU pour les tests quasi-statique et de deux générateurs de pulses 4225-RPM, et d'un autre générateur de pulse HP 8110A pour les mesures en pulsé. Pour les mesures présentées dans la partie suivante, le protocole expérimental a déjà été expliqué dans la partie 2. Pour les mesures dynamiques, nous utiliserons des sondes Tektronix TPP1000 pour sonder la tension aux bornes de l'OxRAM, reliées à un oscilloscope Tektronix DPO5104, doté d'une bande passante de 1 GHz et d'un taux d'échantillonnage en temps réel de 10GS/s.

4.2. Confirmation de la fiabilité du véhicule de test

Avant d'entamer les mesures dynamiques sur des temps courts, nous avons commencé par vérifier que le véhicule de test fabriqué au LETI était bien fonctionnel, et, si oui, comparer ses performances à l'état de l'art, notamment en matière d'endurance. Plus particulièrement nous nous intéresserons à vérifier le bon fonctionnement des dispositifs à 5 plots, puisqu'il s'agit d'une structure très rarement étudiée.

Les dispositifs étudiés ici sont des empilements mémoires de 10nm HfO_2, doté d'une électrode active de titane (cf. Fig. II.11). En effet, comme nous l'avons vu dans le premier chapitre, il s'agit des matériaux les plus connus et les plus étudiés dans la littérature.

Figure II.11 : Représentation schématique de la structure MIM TiN/HfO_2/Ti/TiN étudiée dans cette partie.

Sur le réticule MARS, trois tailles de transistors sont disponibles (notées T1, T3 et T5 avec des W respectifs de W=0.35, 5 et 100µm). Pour les mesures qui suivent, nous avons utilisé des dispositifs 1T1R munis d'un transistor T3 (W=5µm). En Figure II.12, nous avons représenté une caractérisation I-V d'un transistor T3 seul, pour trois différentes tensions de grille. La compliance est, comme nous l'avons dit, fixée par le courant de saturation de chaque Vg. On peut observer que ces transistors sont fonctionnels pour des compliances comprises entre quelques dizaines de µA et un peu plus de 500µA.

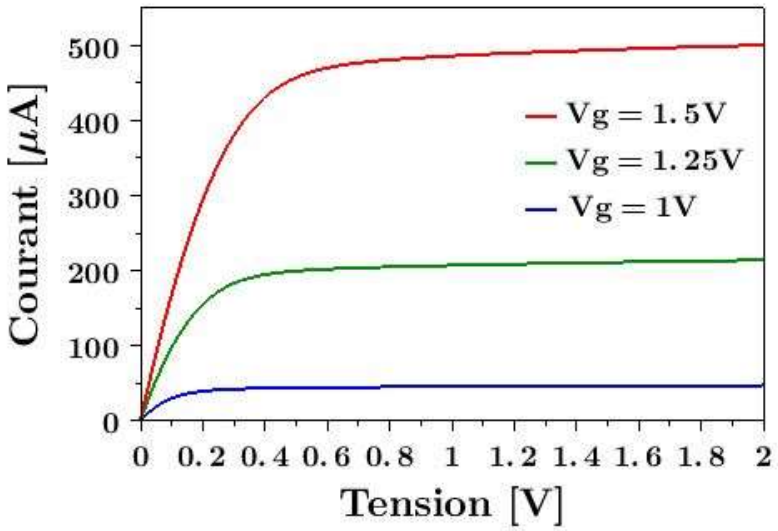


Figure II.12 : Courbe I-V d'un transistor T3, de la même plaque que les OxRAM étudiée, pour des tensions de grilles de 1, 1.25 et 1.5V.

Pour l'instant, le cinquième plot (entre le drain du transistor et l'électrode inférieur de l'OxRAM, cf. Fig. II. 9), n'est pas connecté à la sonde. Les connections sont donc les mêmes qu'avec les dispositifs à 4 plots classiques. Le but est ici de voir si le cinquième plot induit des perturbations, si on utilise les dispositifs de façon classique. En Figure II.13, nous avons tracé les courbes indiquant le courant de compliance obtenu avec chaque transistor, en fonction de la tension de grille. De même que pour la courbe précédente, cette représentation permet de visualiser les domaines de courant sur lesquels chaque transistor contrôle bien le courant. Par exemple le T5 semble opérationnel à partir d'environ 500µA. En dessous, modifier la tension de grille ne permet plus de moduler le courant. De plus nous avons comparé les résultats obtenus avec les dispositifs à 4 et à 5 plots pour voir si le 5^{e} plot modifiait la compliance. On remarque que pour tous les transistors, au-delà d'une tension de grille de 1V, les comportements des deux structures sont presque identiques. Ainsi en ce qui concerne le contrôle du courant le 5^{e} plot, ne gêne pas les mesures, puisque nous n'utilisons quasiment jamais des tensions de grille inférieures à 1V.

Comme nous l'avons dit précédemment, les tests en quasi-statique sont parfaitement adaptés pour vérifier qu'un dispositif commute bien, avec les tensions et les courants espérés. Nous avons donc commencé par réaliser des opérations de set et de reset (après un forming de 3V) en quasi-statique (cf. Fig. II.14).

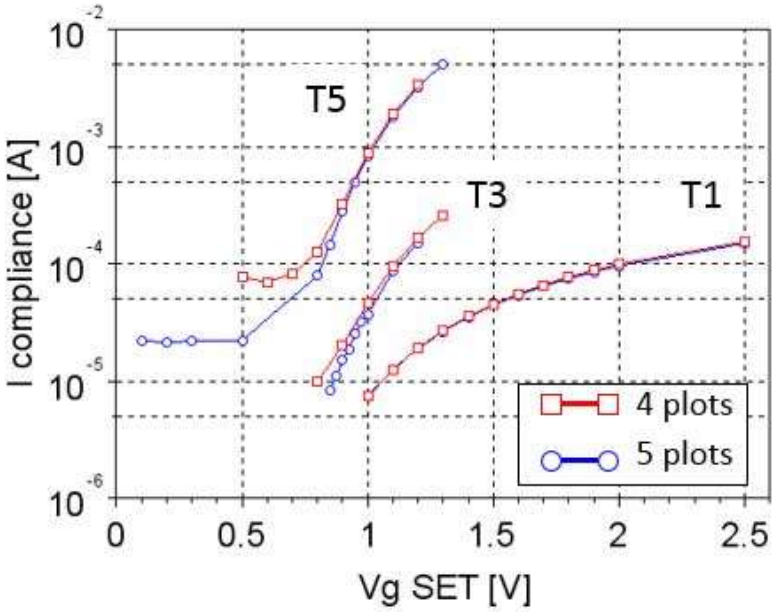


Figure II.13 : Courbe Id(Vg) pour les trois tailles de transistors, et pour les structures avec et sans 5e plot.

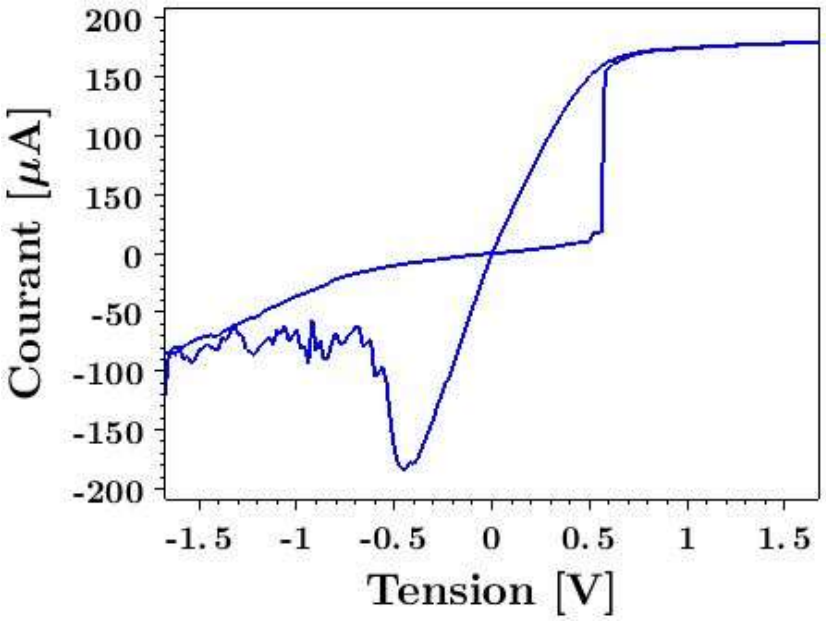


Figure II.14 : Exemple de courbe quasi-statique mesurée sur les dispositifs T3 à 5 plots.

Ces mesures ont été réalisées sous une tension de grille de set de 1.2V (soit une compliance d'environ 180µA) et de reset de 3V. De ce que l'on peut voir sur cette courbe quasi-statique, les dispositifs 5 plots semblent fonctionnels. La tension de set est ici de 0.6V, et celle de reset est de 0.5V. Ces tensions sont assez faibles et témoignent de la capacité des OxRAM à fonctionner à très faibles tensions. On remarque également que le transistor remplit son rôle de compliance, puisque le courant auquel le reset a lieu est à peu près égal au courant

de compliance, ce qui laisse penser qu'il n'y a pas eu d'overshoot de courant durant l'opération de set [6].

Pendant l'opération de reset, comme c'est souvent le cas, on remarque beaucoup de perturbations en courant. En effet, contrairement au set, qui s'arrête au moment où la compliance est atteinte, le reset, continue à mesure que la tension augmente [7, 8]. Deux phénomènes contribuent aux oscillations de courant : à la fois l'augmentation logique du courant à mesure que la tension augmente, et la destruction de portion du filament (via la migration/recombinaison de lacunes d'oxygène, accélérées en température, et en champ électrique) à mesure que le reset se poursuit, qui induit une baisse de la conductivité. Ces oscillations sont donc la signature de phénomènes microscopiques, liés aux ions et lacunes d'oxygène dans l'oxyde d'hafnium.

Après avoir vérifié en quasi-statique le bon fonctionnement des dispositifs OxRAM T3 à cinq plots, nous pouvons passer à leur caractérisation en régime pulsé. Le protocole expérimental utilisé est détaillé dans la partie 2.2. A noter qu'une structure semblable a déjà été testée par Kinoshita et al. dans [5], mais uniquement en régime quasi-statique

Pour ce test d'endurance, nous avons directement voulu démontrer la faisabilité d'écriture et d'effacement sur des temps très courts. En effet pour ces mesures, les pulses de set et de reset étaient des pulses de 20ns. Même si on peut trouver dans la littérature des démonstrations à des pulses encore plus courts [9, 10, 11], 20ns restent des temps extrêmement courts et compétitifs. Il s'agit de la limite imposée par nos générateurs de pulses. Ainsi sur la Figure II.15, nous avons présenté un test d'endurance sur 100 millions de cycles avec ces temps d'écriture et d'effacement de 20ns, sur un dispositif à 5 plots.

Comme nous l'avons dit précédemment, une endurance supérieure à 10^8 cycles est une performance remarquable et proche de l'état de l'art, sur les dispositifs à base d'oxyde d'hafnium [12]. Cela témoigne du très bon fonctionnement de nos véhicules de test et du fait que l'insertion de ce cinquième plot ne gêne pas la fiabilité de nos dispositifs.

De plus, on peut rajouter que ce test a été réalisé à relativement faibles tensions de set (1.5V) et de reset (1.6V). Dans le premier chapitre nous avons abordé le dilemme temps/tension : pour permettre un bon fonctionnement des cellules tout en diminuant le temps de pulse d'écriture et d'effacement, il faut augmenter la tension des opérations [13]. Or ici, nous réalisons un cyclage de 100 millions de cycles, avec des pulses très courts, tout en conservant une faible tension d'opération. On peut cependant remarquer que la fenêtre résistive n'est pas particulièrement bonne. On peut très raisonnablement penser (comme démontré dans [13]) qu'une augmentation de la tension d'opérations permettrait d'améliorer cette fenêtre. Le but de

cette manipulation n'étant pas d'optimiser les performances des dispositifs, mais uniquement de s'assurer de leur bon fonctionnement, nous n'avons pas poussé l'étude en endurance plus loin, afin de nous concentrer sur le cœur de notre étude : l'étude de la dynamique de switching.

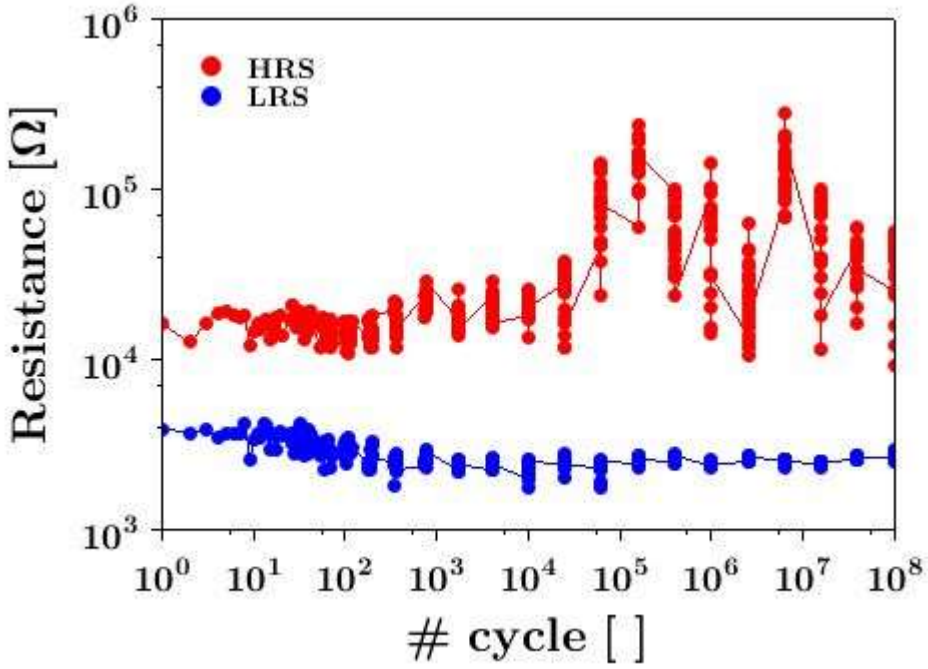


Figure II.15 : Endurance de 100 millions de cycles sur les dispositifs T3 à cinq plots, avec des pulses d'écritures et d'effacement de 20ns.

4.3. Mesure des capacités parasites

Nous pouvons passer à l'étude de la dynamique du switching. Pour ce faire, nous allons maintenant utiliser le cinquième plot dont nous avons parlé précédemment, situé entre le drain du transistor et l'électrode inférieure (cf. Fig. II.9).

Dans la première partie de cette étude, nous nous intéresserons à l'étape de set. Les connections sont schématisées en Figure II.15. Afin de capter les signaux de tensions envoyés sur la grille du transistor et sur l'électrode supérieure, ainsi que la tension résultante sur le cinquième plot, nous avons connecté des sondes Tektronix TPP1000 [14] à un oscilloscope Tektronix DPO5104. Ces sondes présentent une impédance d'entrée de 10MΩ ainsi qu'une capacité de 3.9pF et une bande passante de 1GHz. Comme indiqué sur la Figure II.16, durant le set, la sonde mesure la tension qui retombe aux bornes du transistor. La tension aux bornes

de la mémoire est tout simplement la différence entre la tension appliquée V_{top} et la tension mesurée sur le cinquième plot V_{meas}.

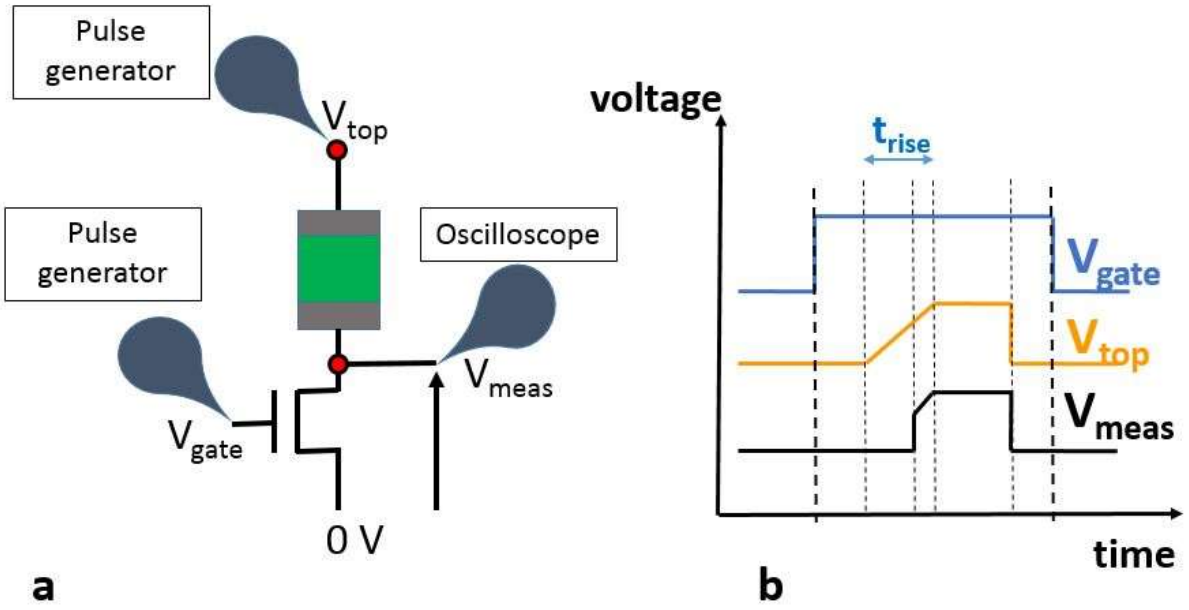


Figure II. 16 : (a) schéma des connections utilisées pendant la mesure de l'opération de set. (b) Schéma des pulses appliqués sur la grille et l'électrode supérieure, et du signal mesuré en V_{meas}.

Comme on peut le voir sur la Figure II.16 (b), le signal envoyé sur la top électrode n'est pas un simple pulse rectangulaire. En effet, le temps de montée de ce pulse, que l'on notera t_{rise}, est réglable et le but sera de faire varier ce temps de pulse sur une large plage d'ordres de grandeur, afin de voir l'évolution de la dynamique de commutation en fonction de cette vitesse de montée.

Durant le set, au moment de la commutation, la résistance de la mémoire est censée chuter brutalement. Ainsi, alors qu'avant la commutation, la mémoire étant bien plus résistive que le transistor, la tension appliquée était très principalement concentrée sur la cellule (donc une tension très faible sur le transistor, en V_{meas}), après celle-ci, la cellule est censée devenir moins résistive (en fonction de la compliance utilisée), et on doit alors mesurer une augmentation de la tension mesurée en V_{meas}, comme indiqué en Figure II.16 (b).

Avant de réaliser les mesures qui nous intéresse nous avons voulu nous assurer que le cinquième plot, ainsi que la sonde qui y est connectée, n'engendrent pas une capacité parasite trop importante, qui pourrait gêner les mesures. En effet, l'ensemble Structure 1T1R + Sonde peut être représenté par le schéma électrique de la Figure II.17.

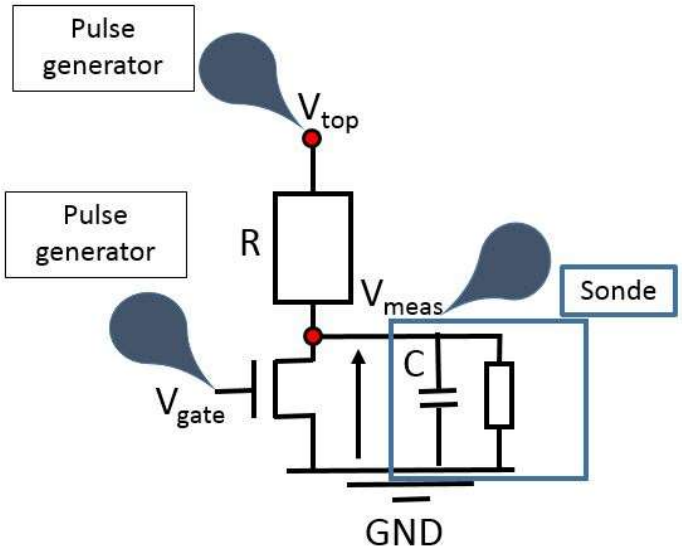


Figure II. 17 : Schéma électrique équivalent de la structure 1T1R+Sonde.

Pour ce faire, nous avons alors modifié légèrement le set-up, en insérant, au plus près du device (c'est-à-dire directement au niveau de la pointe connectée au cinquième plot) un connecteur SMA (Sub Miniature version A). Ensuite, nous avons connecté directement la sonde sur ce connecteur.

a) Set-up classique :

Pour démontrer l'efficacité de ce changement de set-up, nous allons commencer par observer ce qu'il se passe, si on choisit de garder le set-up classique. Pour étudier cette possible capacité parasite, nous avons décidé de réaliser une mesure, durant une opération de set, mais en appliquant une tension nulle sur la grille du transistor. En effet, ainsi, si la sonde ne perturbe pas la mesure, aucun courant n'est censé circuler et aucune commutation ne doit intervenir. Pour ce faire, nous devons avant tout réaliser un forming de la cellule. Pour rappel, un forming est semblable à une opération d'écriture, mais réalisée sur une cellule vierge, très hautement résistive et donc à plus haute tension qu'un set (aux alentours de 3V pour des cellules de HfO_2). Nous avons donc réalisé un forming, tout en connectant la sonde à la cinquième pointe. Nous avons également voulu essayer de réaliser ce forming, sans appliquer de tension sur la grille du transistor. Nous avons mesuré à l'oscilloscope la tension appliquée V_{top} et la tension mesurée V_{meas}. Cette mesure est présentée en Figure II.18. Alors que l'on observe, pendant plusieurs centaines de nanoseconde, une tension V_{meas} nulle, on constate qu'au bout d'un moment (environ 200ns), de façon très brusque, le forming se produit, comme l'indique la montée de V_{meas} (cela traduit une chute de résistance de l'OxRAM). Cela confirme que la capacité parasite

induite par la sonde (et non par l'intégration du cinquième plot en lui-même car lorsque la sonde n'est pas connectée, le dispositif marche parfaitement, comme nous l'avons vu en partie 4.2) crée un chemin pour le courant et court-circuite le transistor. Il s'agit là d'un problème particulièrement important : la structure 1T1R permettant de contrôler le courant perd ici tout son intérêt. On peut d'ailleurs signaler, qu'après ce forming, la cellule est restée bloquée à une très faible résistance (quelques centaines d'Ohm). Cela signifie que l'oxyde a subi un hard breakdown.

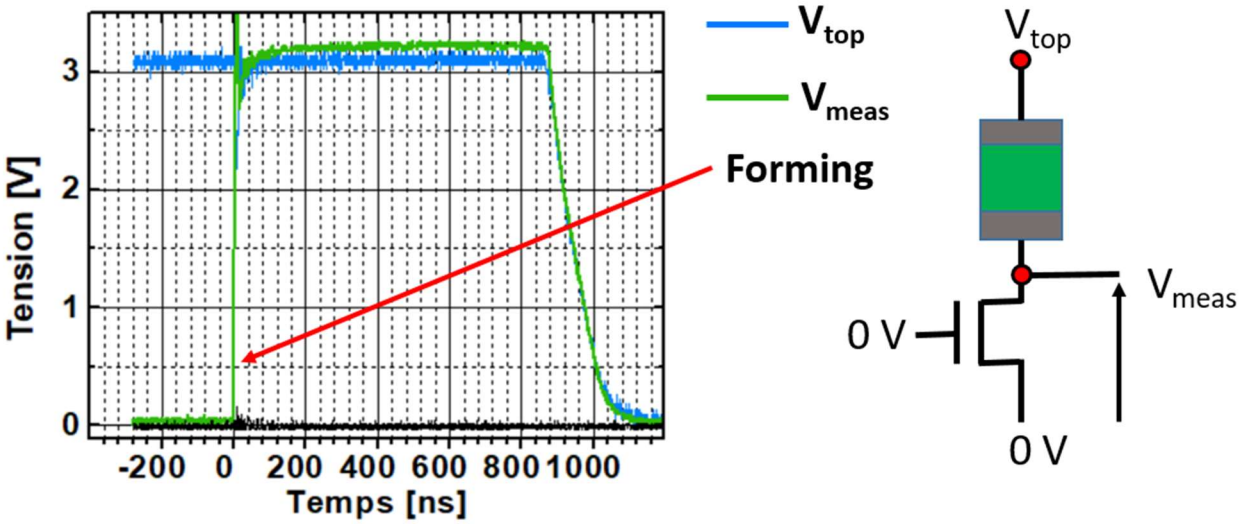


Figure II.18 : Mesure d'une opération de forming réalisée à Vg=0V traduisant le fait que la sonde crée un passage pour le courant et court-circuite le transistor.

Nous avons ensuite utilisé une cellule formée sans connecter la sonde, afin d'induire le hard breakdown. Cette cellule a ensuite été effacée, vers un état HRS de l'ordre de 100kΩ. A partir de cet état, nous avons voulu voir si on observait le même phénomène que celui observé pendant le forming. Nous avons donc connecté la sonde à la cinquième pointe. Puis, nous avons observé l'évolution de V_{meas}, à l'oscilloscope, lorsqu'on applique une tension V_{top} (selon la forme décrite en Fig. II.16 (b)). De même que pour le forming, nous n'avons pas appliqué de tension sur la grille. Normalement, rien n'est censé se passer puisqu'aucun courant ne doit circuler. Or, on observe (cf. Fig. II. 19), comme ci-dessus, que la tension V_{meas} monte encore subitement à un moment, ce qui traduit une écriture de la cellule. Cette mesure confirme encore l'impact de la sonde sur les mesures. De plus on remarque, à la fois avant et après le switching, que V_{meas} met un certain temps à monter, ce qui est caractéristique d'un circuit RC.

Un moyen simple de vérifier que l'on a bien à faire à un circuit de type RC (OxRAM + capacité parasite) est de fitter la courbe représentant V_{meas} au cours du temps de la Figure II.18 avec la charge d'un condensateur au sein d'un circuit RC, soumis à une rampe de tension. La

commutation de l'OxRAM vers le niveau LRS, se traduira par le changement de la valeur de R. Le fit de cette courbe est représenté en Figure II.20.

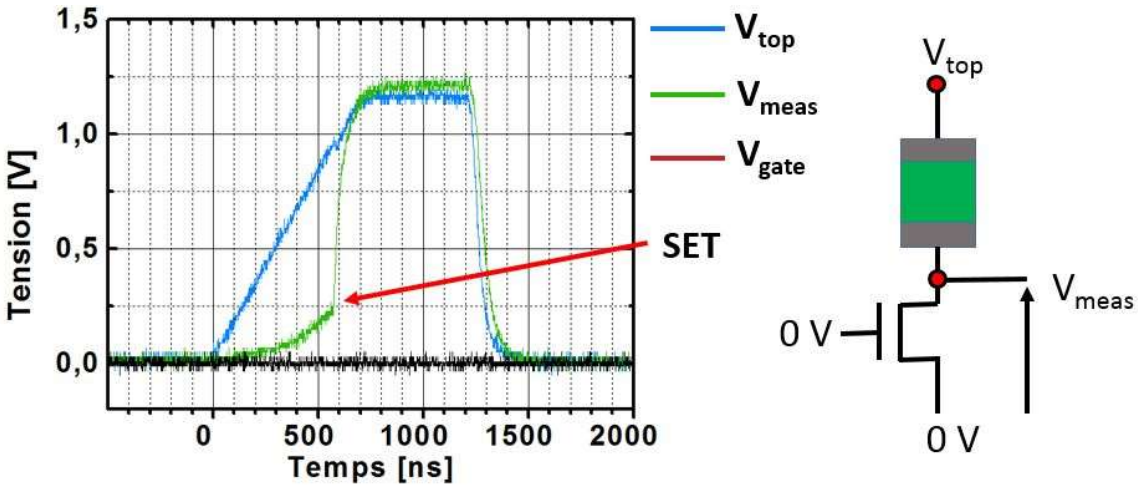


Figure II.19 : Mesure d'une opération de set réalisée à Vg=0V.

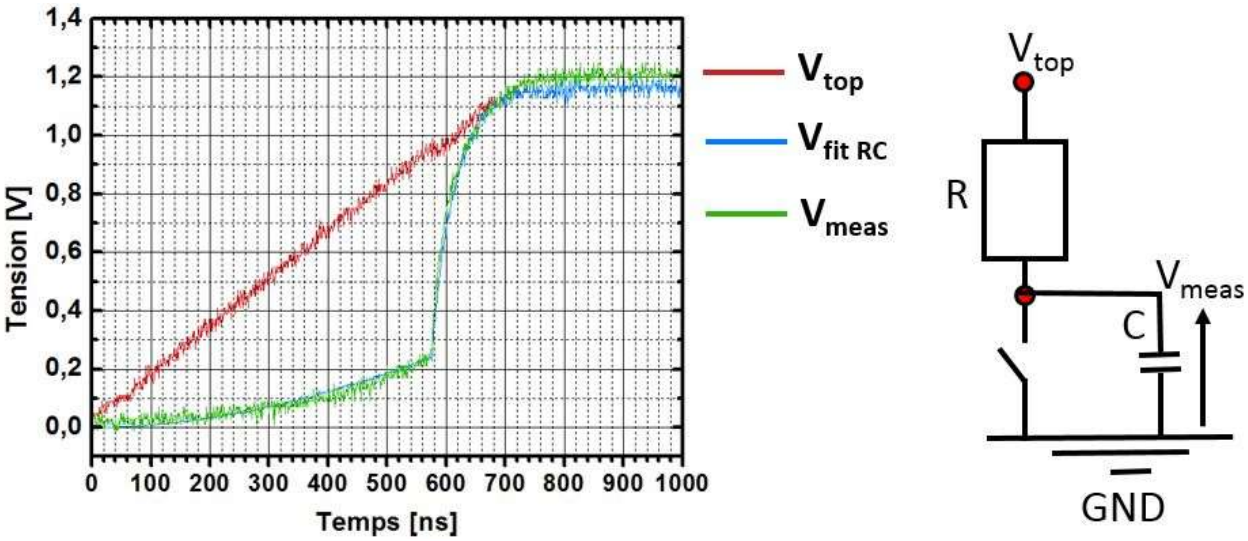


Figure II.20 : Fit de V_{meas} par la charge d'un condensateur soumis à une rampe de tension. La commutation est modélisée par le changement de la valeur de R. Le transistor, dont la tension appliquée sur la grille est nulle, est modélisé par un interrupteur ouvert.

Le tracé de $V_{fit\ RC}$ est simplement le tracé des solutions des équations électroniques. Les paramètres utilisés pour ce fit sont :

- La résistance avant la commutation R_{HRS}=100kΩ
- La résistance après la commutation R_{LRS}=1315Ω
- La capacité du condensateur (donc de la sonde) C=20pF

Ainsi, lors de la commutation vers l'état faiblement résistif, la constant de temps τ=RC passe de 2µs à 26.3ns.

On constate que la modélisation par un simple circuit RC, avec un changement de résistance, permet de fitter efficacement V_{meas}. Cela confirme encore une fois que le problème vient de la capacité parasite induite par la sonde. En effet, la valeur de 20pF pour la capacité trouvée ici est bien supérieure aux 3.9pF renseignée sur la fiche technique de la sonde. Cette capacité parasite provient donc probablement des connections, liées à l'ajout de la sonde et non à la sonde elle-même. Il devient alors impératif de trouver un moyen d'abaisser fortement cette capacité. En effet, une constante de temps de 26ns est trop élevée dans notre cas, puisque l'objectif est de capter la dynamique de la commutation sur des temps inférieurs à 100ns, ce qui n'est pas possible avec une constante de temps si élevée.

Ainsi, sans modification préalable du set-up nous obtenons une capacité parasite de 20pF, bien trop importante pour les mesures que nous voulons réaliser.

b) Utilisation d'un connecteur SMA pour placer la sonde au plus près du device

Nous pouvons alors démontrer qu'avec les modifications de set-up, nous améliorons considérablement les capacités parasites. Pour ce faire, nous avons réalisé une opération de reset sur une cellule préparée dans un état LRS. Nous avons ici choisi de réaliser cette mesure pendant le reset, pour avoir volontairement une constante RC plus élevée, nous permettant de la mesurer plus précisément à l'oscilloscope. Puis nous avons mesuré la tension V_{meas}, au moment de la fin du pulse de reset. Pendant le reset, comme l'électrode supérieure est connectée à la masse, V_{meas} est maintenant la tension aux bornes de l'OxRAM. A la fin du pulse, V_{meas} revient à 0V avec une constante de temps τ=RC que l'on peut déterminer grâce à la mesure de V_{meas} (cf. Fig. II.21).

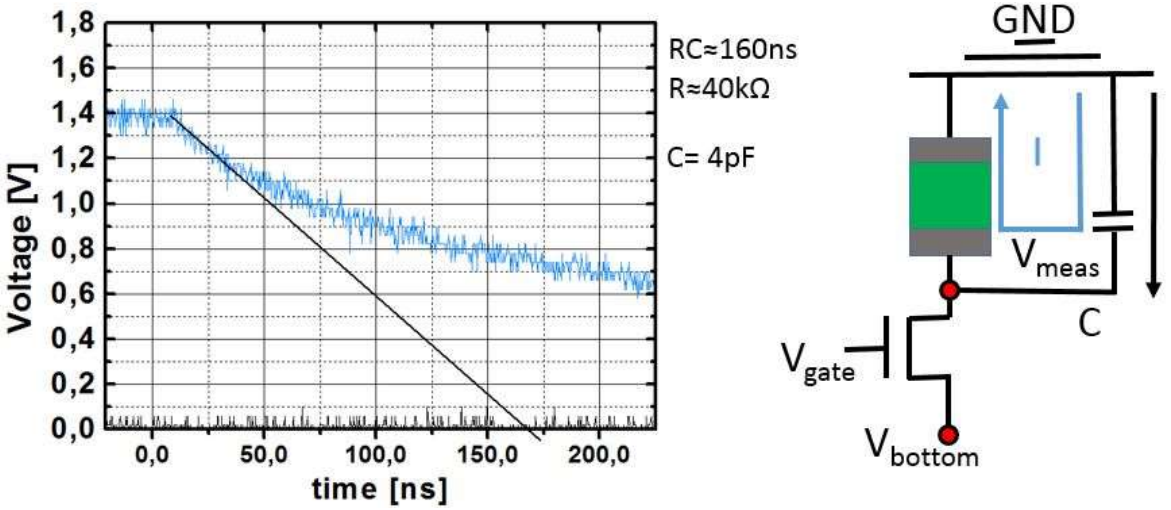


Figure II.21 : Mesure de V_{meas} à la fin du pulse de reset, assimilable à la décharge d'un condensateur. La constante de temps est déterminée très facilement par la tangente à la courbe, au début de la décharge.

On mesure ainsi une constante de temps de 160ns. Par la suite, via une opération de read, on lit la résistance de la cellule et on obtient une valeur de 40kΩ. On en déduit donc une valeur de capacité de 4pF. Cette valeur est bien inférieure à celle de 20pF que l'on avait avant la modification du set-up. Par ailleurs, on remarque qu'elle correspond presque parfaitement à la valeur de 3.9pF de la datasheet de la sonde. Cette valeur peut être comparée à d'autres études présentes dans la littérature, notamment [15], où Ielmini et al. réalisent une étude en dynamique du comportement de mémoires en NiO, avec une capacité parasite mesurée à 16pF.

De plus, le problème des forming et écriture alors que la tension appliquée sur la grille du transistor était nulle, a disparu. Maintenant que nous avons réussi à éliminer ce problème et à diminuer de façon importante la capacité parasite, nous pouvons passer à la visualisation « en direct » du phénomène de commutation.

4.4. Observation directe du set

On peut maintenant revenir aux connections schématisées précédemment en Figure II.15, pour l'étude de la dynamique du set. En Figure II.22, on peut voir un exemple de mesure observée à l'oscilloscope. Ici, la tension de set appliquée est de 1.4V et la vitesse de la rampe sur V_{top} est de 200ns. L'évènement de commutation est très facilement visible, avec la très soudaine montée de la tension aux bornes du MOS (V_{meas}). Avant le switching, la cellule étant en HRS, très peu de tension retombe sur le transistor. Avec le set et la chute de résistance de l'OxRAM, la tension aux bornes du MOS augmente brutalement. Après chaque mesure, une lecture de la résistance est réalisée afin de s'assurer que le set a bien eu lieu et que l'on n'a pas mesuré un quelconque artefact de mesure.

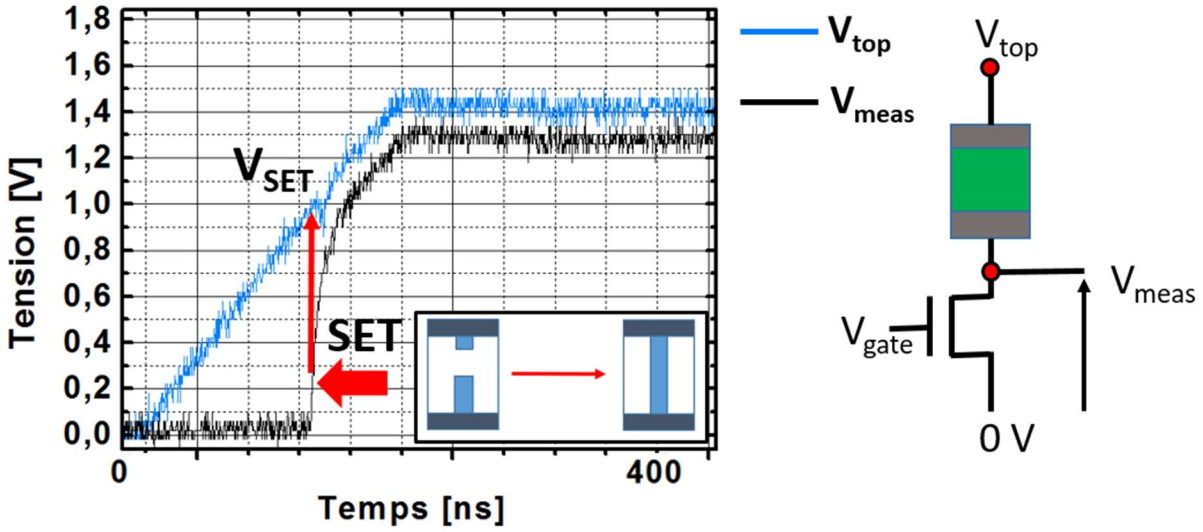


Figure II.22 : Capture de la commutation vers l'état LRS à l'oscilloscope, lors d'une rampe de tension de 200ns appliquée sur l'électrode supérieure. Nous avons représenté en inset l'évolution du filament conducteur lors de l'opération : le set signifie la reconstruction du filament.

Nous pouvons également étudier la chute de tension liée à la présence du transistor, avant et après le set. Pour ce faire, il suffit de tracer les courbes de point de fonctionnement entre une mémoire en état LRS (ici, les résistances LRS sont de l'ordre de 1500Ω) ou HRS (100000Ω) et un transistor MOS T3. Cette visualisation des points de fonctionnement, pour une tension appliquée de 1.5V est représentée en Figure II.22. Ici, pour simplifier le problème nous avons simplement schématisé les OxRAM comme des résistances. Cette courbe se lit de la façon suivante : si le switching (d'un état de 100000Ω vers un état de 1500Ω) apparait à une tension de 1.5V, la tension aux bornes du MOS va passer de quelques millivolts à environ 1.2V (les deux tensions correspondantes aux deux points de fonctionnement). On constate bien qu'en état HRS, pratiquement aucune tension ne retombe sur le transistor, tandis qu'une majeure partie de la tension retombe sur le MOS, une fois l'opération de SET réalisée.

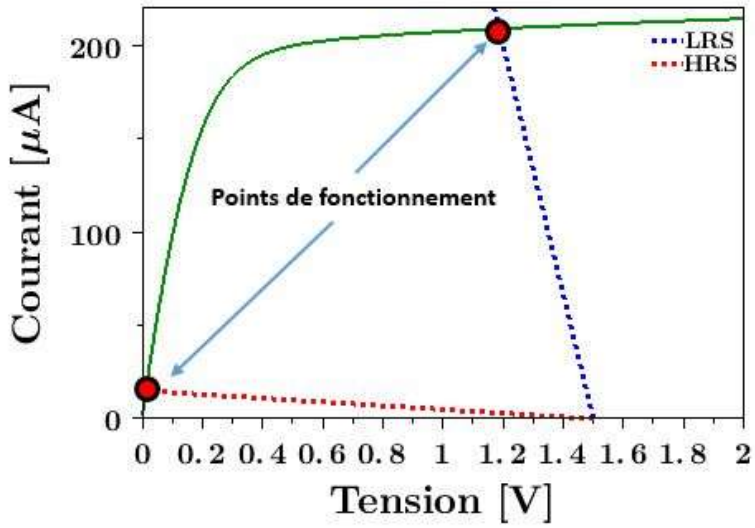


Figure II.23 : Points de fonctionnement avant et après le set, pour une tension de 1.5V appliquée sur l'électrode supérieure. Ici la tension représente la tension aux bornes du MOS.

Comme on pouvait s'y attendre au vu de la littérature [16, 17], le switching est, pour l'étape de set très abrupt. Grâce à la résolution de 400ps permise par l'oscilloscope utilisé, nous sommes capables de détecter avec précision un tel évènement. Nos générateurs de pulses nous permettant d'obtenir des temps de montée de 20ns, nous avons réalisé la même mesure sur une rampe sur V_{top} de 20ns (cf. Fig. II.24).

Sur cette figure, on observe très clairement l'évènement de set, en environ 15ns, à une tension de 1.2V. Le fait d'obtenir un temps de commutation si court, et d'arriver à le détecter mérite d'être signalé. Comme nous l'avons dit, très peu d'étude démontre l'observation en dynamique de telles performances.

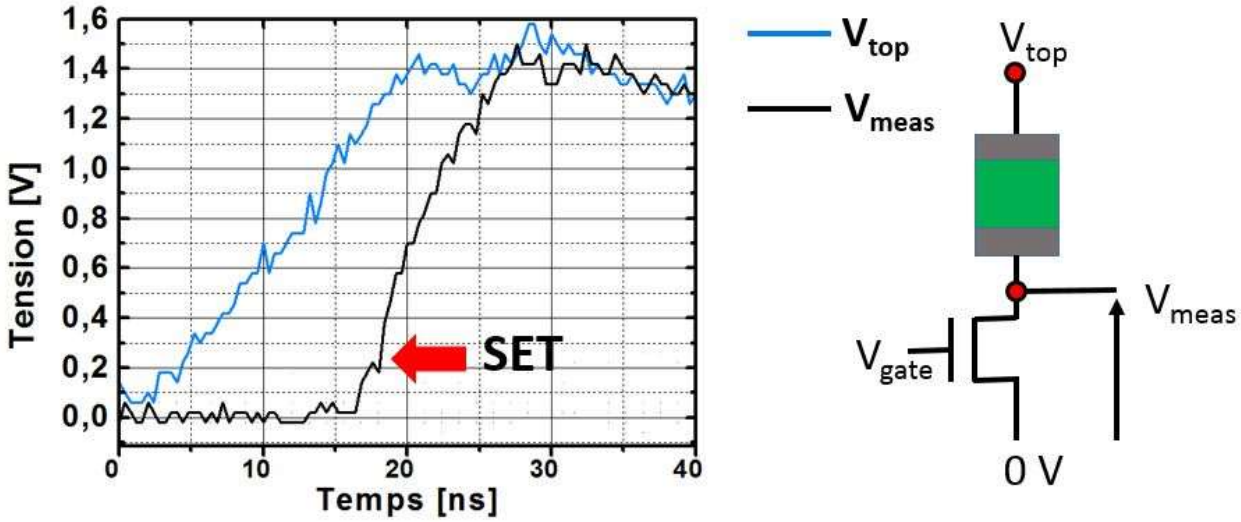


Figure II.24 : Capture de la commutation vers l'état LRS à l'oscilloscope, lors d'une rampe de tension de 20ns appliquée sur l'électrode supérieure.

Ainsi, l'un des grands intérêts de cette manipulation est d'extraire avec une grande précision la tension réelle à laquelle le set se produit. En effet, même si on applique un pulse de 1.4V, on constate que la commutation se produit avant, durant la rampe de montée. Comme indiqué sur la Figure II.22, nous désignerons par V_{SET} la tension appliquée sur l'électrode supérieure au moment où on observe la transition. Or, cette mesure nous permettant de faire varier la vitesse de la rampe sur une grande plage d'ordres de grandeur, nous avons décidé d'étudier l'impact de cette vitesse de rampe sur la valeur V_{SET} de la tension de set.

4.5. Influence de la vitesse de rampe sur la tension de set

Pour s'assurer que nos mesures sont comparables, d'une rampe à une autre, toutes les cellules sont, avant la mesure de la tension de set, préparées dans un état HRS aux alentours de 100kΩ. En effet, une rapide étude préliminaire nous a permis de voir que la résistance initiale dans laquelle se situe une cellule joue un rôle sur la tension de set, comme l'indique la Figure II.23. En effet plus la résistance initiale est élevée, plus le filament est rompu sur une grande distance. On peut donc prévoir qu'une plus grande tension sera nécessaire pour reconstruire ce filament. Pour cette figure, nous avons réalisé 40 mesures de la tension de set, pour différentes résistances initiales. On observe que celle-ci joue un rôle assez important dans la tension de set. Sur l'échantillon testé on remarque également qu'en dessous de 100kΩ, la tension V_{SET} peut varier de façon importante. C'est la raison pour laquelle nous avons décidé de préparer toutes les cellules à 100 kΩ, pour les mesures de V_{SET} de la manipulation suivante.

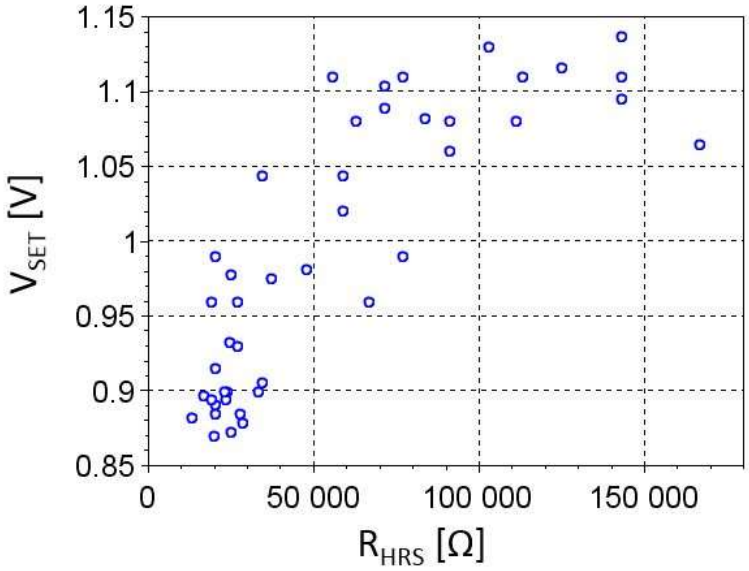


Figure II.25 : Impact de l'état HRS avant le set, sur la tension de set.

Nous avons donc répété les manipulations des Figure II.22 et 24, pour différentes valeurs de vitesse de rampe β. Nous avons fait varier cette vitesse entre $1{,}5.10^3$ et $7{,}5.10^7$ V/s. Le résultat de l'impact de β sur la valeur de V_{SET} est représenté en Figure II.26. Sur cette courbe, chaque point représente 30 mesures de valeurs de V_{SET}. Comme C. Cagli et al. l'ont remarqué dans [18] (dans des structures 1R en revanche), cette courbe peut être fittée, en échelle logarithmique par une droite, du type : $V_{set} = a.log_{10}\left(\beta \Big/ \frac{1V}{s}\right) + b$ Ainsi l'influence de la vitesse de rampe sur la tension de set est logarithmique. Augmenter la vitesse de la rampe de plusieurs ordres de grandeur augmente la tension à laquelle le Set se produit. Cela s'explique par le fait qu'augmenter la vitesse de rampe diminue la durée sur laquelle la mémoire est exposée à un champ électrique, donc, logiquement, réduit la probabilité que la cellule switche.

Cette courbe confirme l'idée, très répandue dans la littérature [7, 8, 19] que la génération de lacunes d'oxygène responsable de la création du filament conducteur est activée exponentiellement par le champ électrique. Une légère augmentation de la tension peut compenser des différences en termes d'ordre de grandeur de la vitesse de rampe. Le set est donc bien une opération commandée en tension. Du point de vue des performances cela est très intéressant puisque cela signifie que les OxRAM peuvent atteindre des temps de switching très rapides, au prix d'une augmentation relativement légère de tension mise en jeu.

A noter qu'il s'agit de la première fois que cette dépendance de la tension de set avec la vitesse de rampe est montrée sur une structure 1T1R. Ces travaux ont fait l'objet d'une

présentation orale lors de la conférence IIRW (International Integrated Reliability Workshop) 2015, en Californie [20].

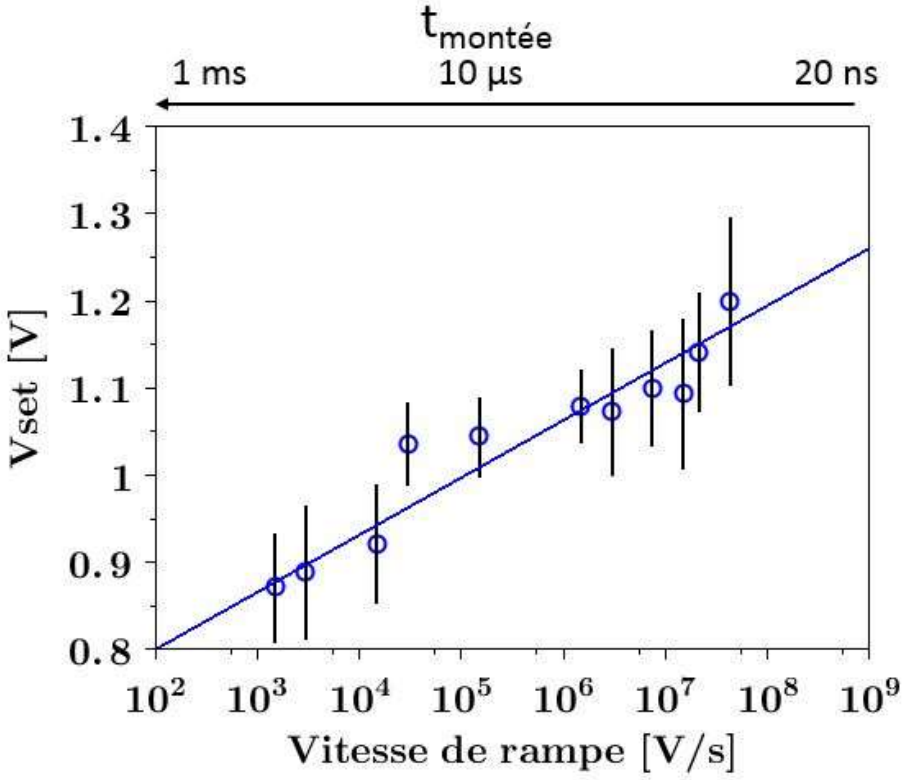


Figure 26 : Impact de la vitesse de rampe β sur la tension de set. La courbe bleu est un fit linéaire.

4.6. Mesure en température

Si l'impact en champ électrique vient d'être démontré, il est également connu que les phénomènes de breakdown [8, 21] ou de migration [7, 22] sont également activés en température, selon une loi d'Arrhenius. Beaucoup de papiers étudient aussi le chauffage lié au passage du courant dans les OxRAM [23, 24, 25]. C'est pourquoi les mêmes mesures ont été réalisées à d'autres températures. De plus, de telles mesures permettent de confirmer la stabilité des dispositifs en température, ce qui est également intéressant, du point de vue de la fiabilité.

En effet le banc de test est muni d'un chuck chauffant, qui permet de mettre les dispositifs à la température désirée et de réaliser les mesures à ces températures. Nous avons donc réalisé ces manipulations entre 25°C (Fig. II.26) et 250°C. Comme l'opération de set est censée être activée en température, on s'attend à une baisse des valeurs des tensions de set lorsque la température augmente.

La courbe correspondante à l'impact de la vitesse de rampe β sur la tension de set pour différentes température est représentée en Figure II.27. Comme prévu on observe un abaissement de la tension de set lorsque la température augmente. Ces mesures ont été effectuées à 25, 100, 150, 200 et 250°C, mais seules les mesures à 25, 150 et 250°C ont été

représentée, pour un souci de lisibilité. La tension de set diminue d'environ 0.1V lorsque la température passe de 25°C à 250°C, soit une variation d'environ 0.04%/K, ce qui est relativement faible.

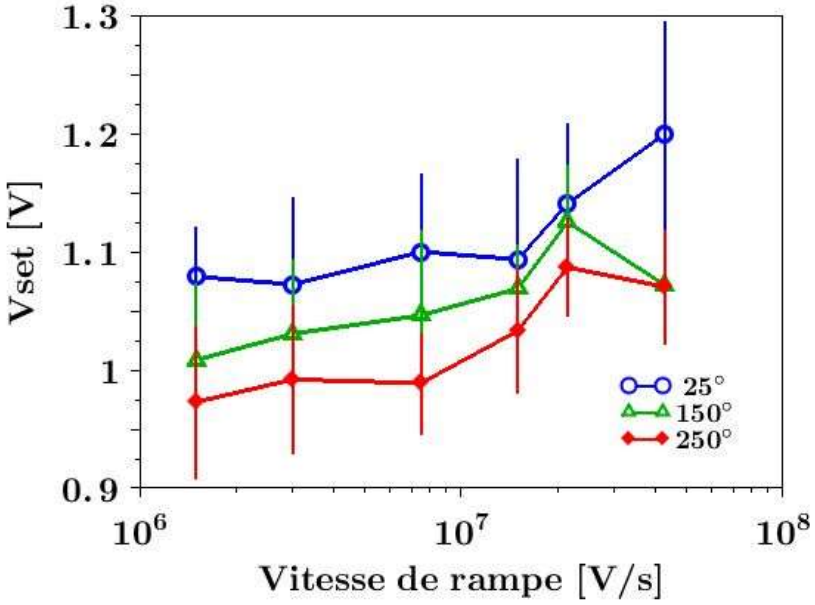


Figure II.27 : Impact de la vitesse de rampe β sur la tension de set pour 3 températures différentes.

On retrouve bien l'impact de la vitesse de rampe sur la tension de set : les trois courbes sont parallèles.

Maintenant que ces mesures ont été réalisées pour l'étape d'écriture, nous pouvons passer à l'étude en dynamique de l'étape d'effacement.

4.7. Cas du reset

Les mesures en dynamique du reset se sont avérées plus délicates. En effet, ces mesures diffèrent de celles sur l'opération de set sur plusieurs points, comme nous allons le montrer dans cette partie.

4.7.1 Différentes connections

Lors du reset, une tension positive est appliquée sur l'électrode inférieure, tandis que l'électrode supérieure est connectée à la masse (cf. Fig. II.28).

Ainsi, durant le reset, la sonde connectée entre l'OxRAM et le drain du transistor ne mesure plus la tension aux bornes de ce dernier, mais la tension aux bornes de la mémoire. C'est pourquoi, comme indiqué en (b), on doit encore observer ici une augmentation de la tension V_{meas} lorsque le reset survient (la résistance de l'OxRAM augmente lors du reset, donc la tension à ses bornes également).

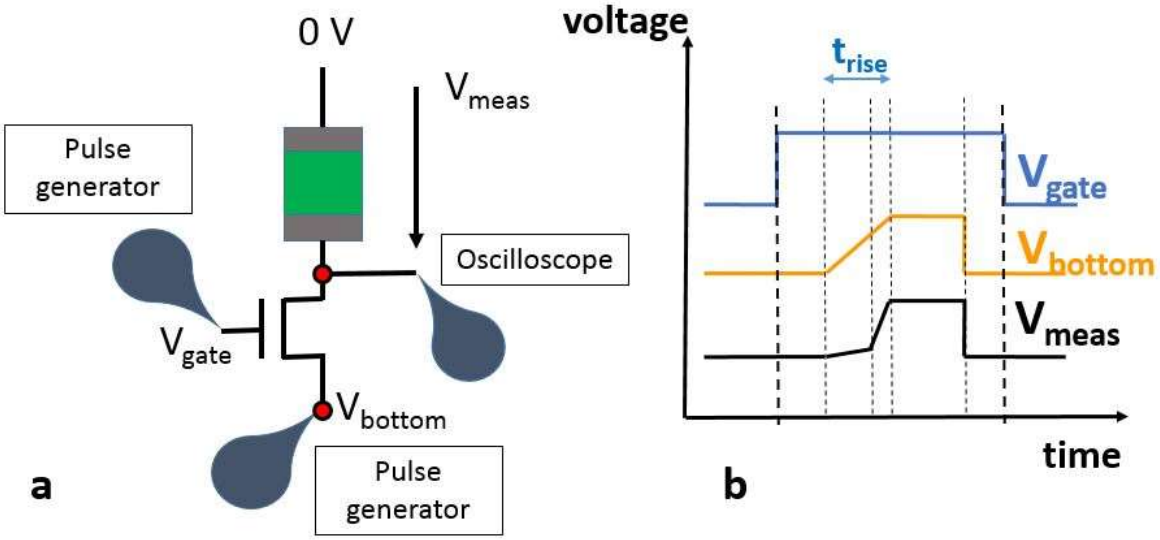


Figure II.28 : (a) schéma des connections utilisées pendant la mesure de l'opération de reset. (b) Schéma des pulses appliqués sur la grille et l'électrode inférieure, et du signal mesuré en V_{meas}.

4.7.2 Polarisation du transistor

Lors du set, la source du transistor est constamment connectée à la masse. Ainsi, seule sa tension de drain est en mesure d'augmenter. Les courbes Id-Vd classiques, à Vs=0V, de caractérisation du MOS sont donc complètement valables pour décrire le point de fonctionnement du 1T1R. En revanche, lors du reset ni la source ni le drain du transistor n'est connecté à la masse. On ne peut donc pas utiliser les courbes Id-Vd classiques pour prévoir le point de fonctionnement. Nous avons donc du caractériser nos transistors d'une autre manière, schématisée en Figure II.29 (a). Un exemple de courbe de caractérisation d'un transistor T3 de cette manière est représenté en Figure II.29 (b), pour une tension de bottom de 1.8V.

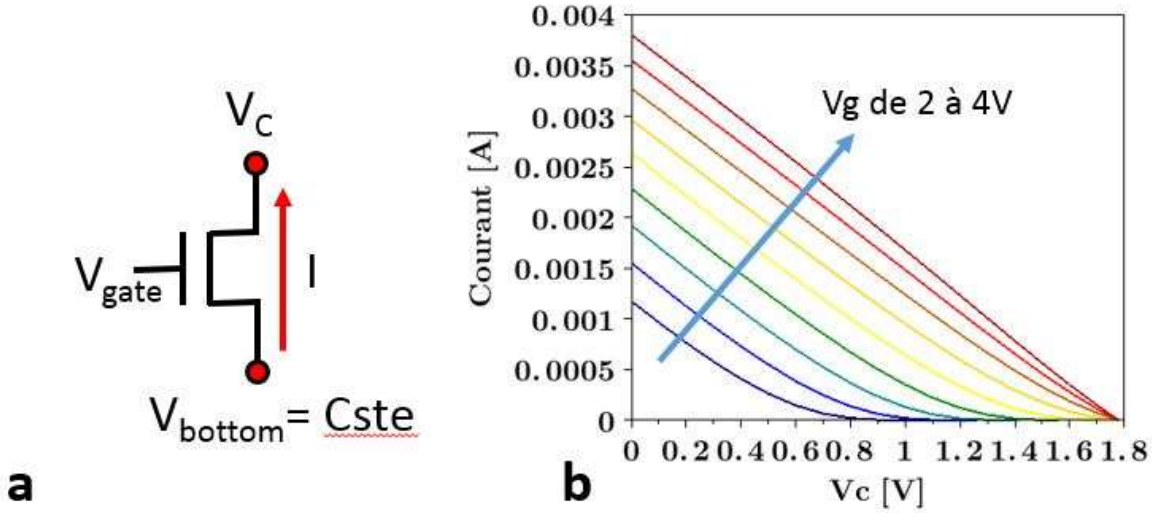


Figure II.29 (a) Polarisation du transistor pour les caractérisations. La tension de bottom est laissée constante, tandis que V_C varie de 0 à V_{bottom}. (b) Caractérisation d'un transistor avec ces connections. Ici V_{bottom}=1.8V.

A mesure que la tension V_C se rapproche de V_{bottom}, le courant diminue, puisque la tension aux bornes du transistor diminue.

Comme V_C symbolise au final la tension aux bornes de l'OxRAM on peut maintenant déterminer le point de fonctionnement du circuit (cf. Fig. II.30).

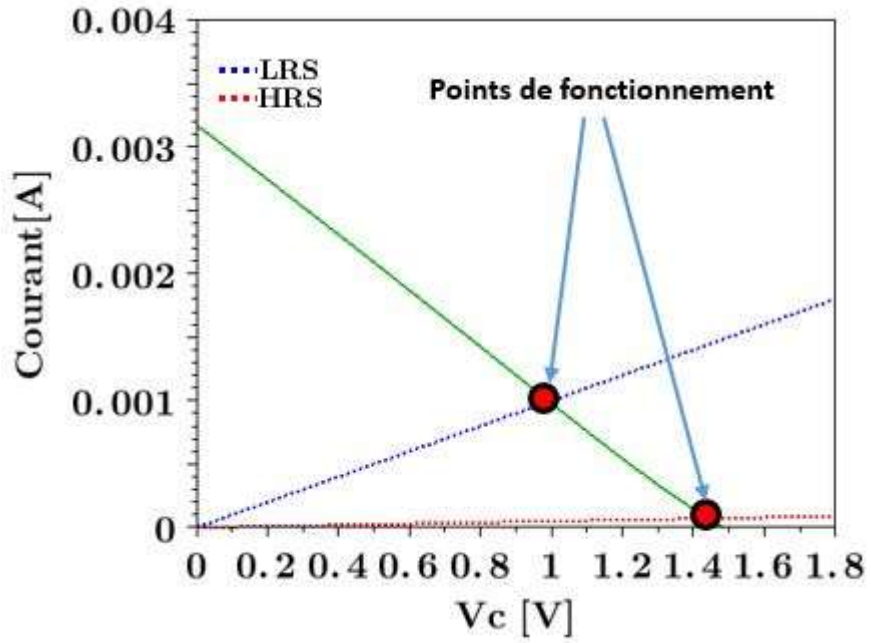


Figure II.30 : Points de fonctionnement avant et après le reset, pour une tension de 1.5V appliquée sur l'électrode inférieur. Ici la tension représente la tension aux bornes de la mémoire.

Cette courbe se lit de la façon suivante : si le switching (d'un état de 1000Ω vers un état de 20000Ω ici) apparait à une tension de 1.5V, la tension aux bornes du MOS va passer d'environ 1V à environ 1.4V (les deux tensions correspondantes aux deux points de fonctionnement). Ainsi l'amplitude du changement de tension est bien inférieure à celle observée lors d'un set, ce qui rend sa détection plus compliquée. C'est d'autant plus vrai si l'état LRS n'est pas aussi faiblement résistif que celui représenté sur la courbe. En effet, avec un état LRS de 4000Ω, le switching n'est presque plus détectable (différence de 0.1V entre les deux états) (cf. Fig. II.31). Ainsi, pour que nos reset soient détectables nous devons préparer nos cellules dans des états LRS très faiblement résistifs, de l'ordre du kΩ.

Enfin, comme nous l'avons montré dans le premier chapitre sur l'état de l'art, le reset est connu pour être une opération progressive (au contraire du set qui est une opération abrupte) [7, 8, 26]. Ainsi, alors que pour le set, on sait que lorsque celui-ci commence, on va observer un phénomène d'une durée très courte : dès le moment d'observation de la commutation, la résistance de la cellule chute très rapidement. Au contraire, lorsque le reset commence, la résistance augmente, mais plus progressivement. Il n'y a donc pas de moment où la résistance

passe instantanément d'une valeur faible à une valeur élevée. Ce point rend également la détection du reset plus délicate.

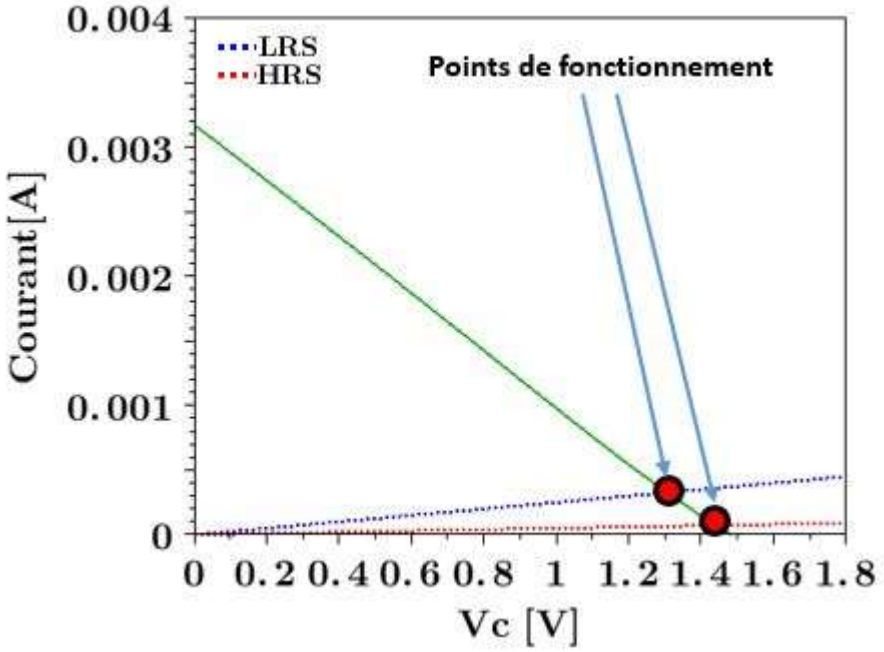


Figure II.31 : Points de fonctionnement avant et après le reset, pour une tension de 1.5V appliquée sur l'électrode inférieur, pour un LRS de 4000Ω.

4.7.3 Détection du reset

Les points de différence avec l'opération de set ayant été éclaircis, nous pouvons passer à l'observation du phénomène de reset. Comme nous l'avons dit précédemment, les OxRAM sont préparées dans un état LRS très faiblement résistif de résistance inférieure à 1kΩ.

En Figure II.32, nous avons représenté une mesure d'un reset observée à l'oscilloscope, ainsi que les connections associées. Ici, la tension appliquée à l'électrode inférieure est programmée pour atteindre une tension de 1.8V en 250ns.

Contrairement au set, avant le switching, toute la tension V_{bottom} ne retombe pas que sur la mémoire, mais nous avons un diviseur de tension. La tension aux bornes de l'OxRAM est la tension V_{meas}, c'est pourquoi, la tension V_{RESET} (définie comme la tension aux bornes de la mémoire au moment du début du reset), n'est pas mesurée sur la voie V_{bottom}, mais sur la voie V_{meas}, comme représentée sur la Figure II.32.

On peut remarquer que, comme prévu, le switching est moins facilement détectable que pour l'étape de set.

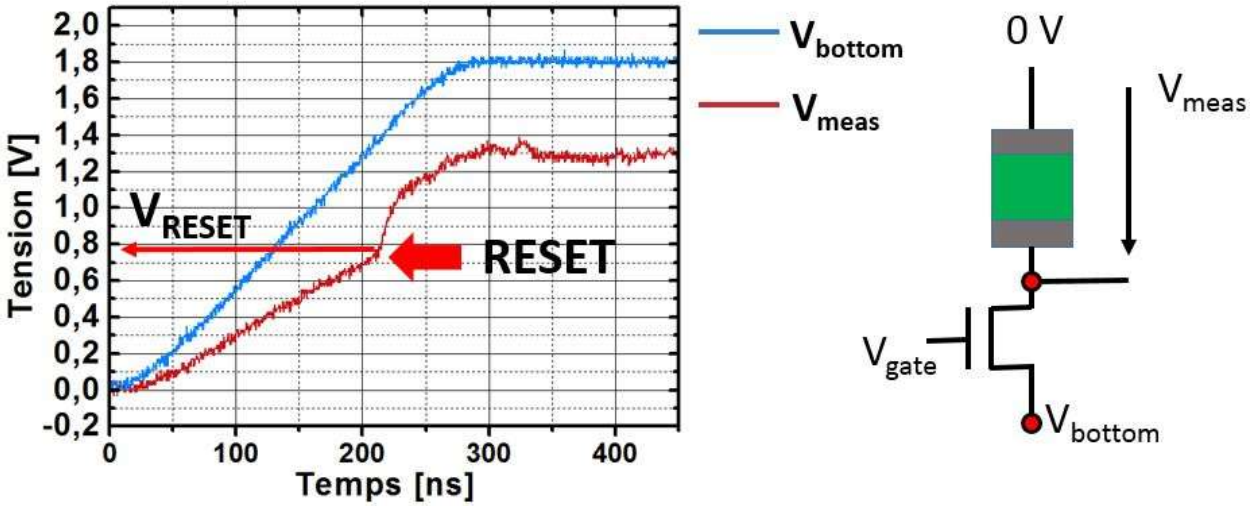


Figure II.32 : Capture de la commutation vers l'état HRS à l'oscilloscope, lors d'une rampe de tension de 250ns appliquée sur l'électrode inférieure.

4.7.4 Influence de la vitesse de rampe sur la tension de reset

De même que pour le set, nous avons mesuré la tension V_{RESET}, sur une plage importante de vitesse de rampe β, entre $1{,}5.10^3$ et $7{,}5.10^7$ V/s. La courbe obtenue est représentée en Figure II.33. L'allure de la courbe est très similaire à celle obtenue lors du set. On peut la fitter par une droite ce qui témoigne d'un impact logarithmique de la vitesse de rampe sur la tension de reset. Là encore, augmenter la vitesse de rampe de plusieurs ordres de grandeur augmente la tension de reset de quelques centaines de mV.

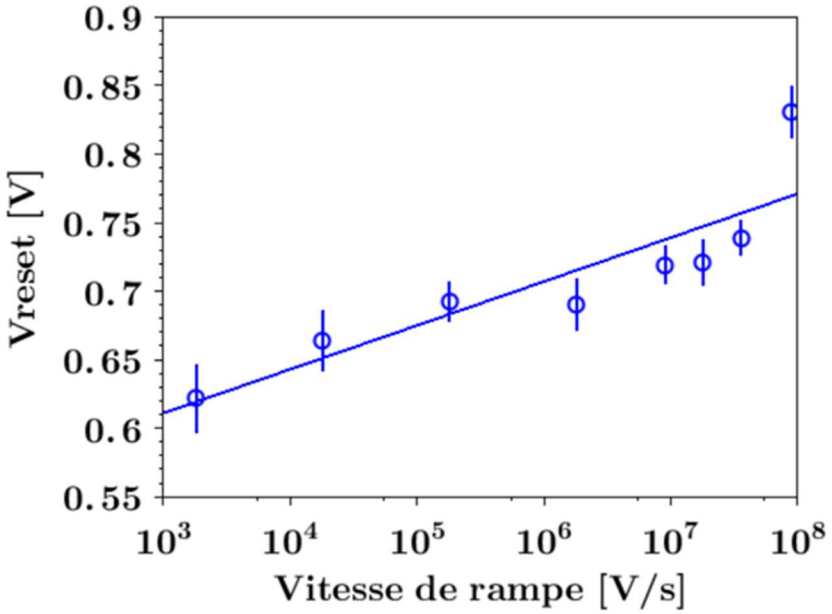


Figure II.33 : Impact de la vitesse de rampe β sur la tension de reset.

Dans cette partie nous avons démontré des mesures en dynamique sur le switching en écriture et en effacement de cellules mémoires OxRAM, jusqu'à des temps de l'ordre de la dizaine de nanoseconde. Les dispositifs semblent fonctionner sur des temps très courts. C'est-à-dire qu'ils sont capables de commuter d'un état faiblement (ou fortement) résistif vers un état fortement (ou faiblement) résistif. Néanmoins, comme nous l'avons vu dans le premier chapitre, baisser les temps de pulse se fait parfois au détriment d'une variabilité plus élevée [27, 28]. C'est le point que nous avons décidé d'éclaircir dans la partie suivante.

5. Impact sur les distributions

Le but de cette partie est d'analyser l'impact de la réduction du temps de pulse utilisé pour switcher les OxRAM sur leur variabilité, c'est-à-dire sur la dispersion des distributions des différents états résistifs. Pour cette étude, nous avons décidé de faire également varier les tensions et les courants utilisés, afin d'obtenir une analyse multiparamétrique complète. Tout d'abord nous n'aborderons que l'étape de set, afin d'être le plus clair possible.

5.1. Protocole de mesures

Pour ces mesures, le banc de test utilisé est le même que pour les tests précédents. Nous continuerons à utiliser le mode pulsé.

Ici nous n'allons plus avoir besoin de faire varier la rampe des pulses de set. Celle-ci sera laissée fixe à 20ns (la plus faible valeur possible). Ici, ce qui varie, c'est la durée du plateau du pulse (cf. Figure II.34).

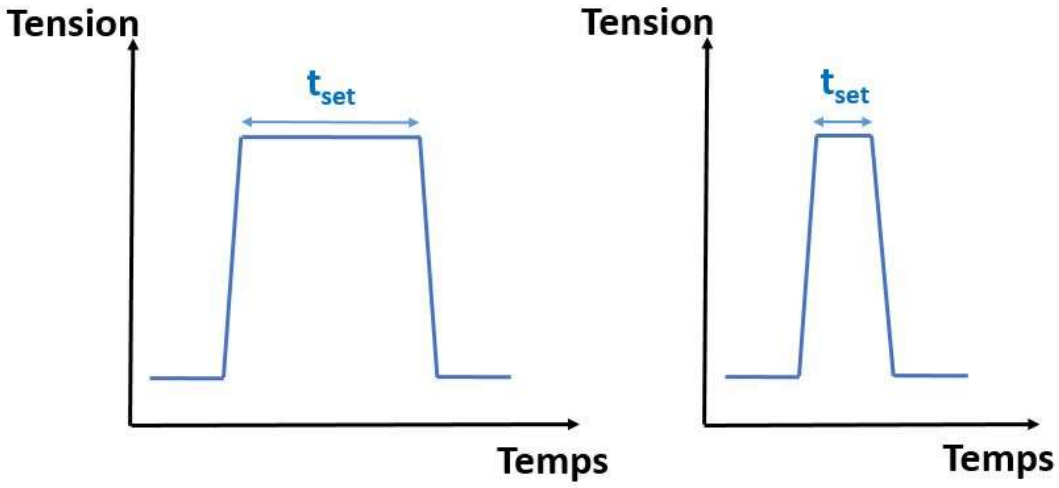


Figure II.34 : Lors de ces manipulations, nous allons faire varier la durée du plateau, et non la vitesse de rampe.

Le but de ces manipulations sera de réaliser des mesures d'endurances de 400 cycles d'écriture-effacement (400 sets et 400 resets). De telles mesures seront réalisées sur plusieurs dispositifs afin de comparer les variabilités cycle-à-cycle et device-à-device. Ainsi nous obtiendrons des courbes de distribution, comme présentées dans la partie 2.3.

Les trois paramètres que nous pouvons faire varier, dans le cadre de l'opération de set sont le courant de compliance Icc (commandée par la tension de grille sur le transistor), la durée du plateau de pulse t_{SET} et la tension appliquée sur l'électrode supérieure V_{top}. Comme nous l'avons vu précédemment, le set apparaît en général très vite, avant même que la tension appliquée sur l'électrode supérieure n'atteigne sa valeur finale. Une fois le set déclenché, une grosse partie de la tension retombe sur le transistor. On peut donc supposer que la tension appliquée sur l'électrode supérieure n'aura pas une influence capitale (comme le set dépend de façon exponentielle du champ électrique, lorsque la tension est trop faible, le champ n'a vraiment quasiment aucun impact sur le filament). C'est pourquoi nous avons décidé de commencer par étudier ce paramètre.

5.2. Influence de la tension de set sur les distributions LRS

Pour la première partie de cette étude, nous avons testé 12 dispositifs, tous de type T3 (c'est-à-dire avec un transistor d'un W de 5µm).

Ces dispositifs ont été testés pour différents courants de compliance (200, 300 et 400µA), avec un temps de pulse t_{SET} de 2µs. Nous avons testé 3 tensions de set différentes : 1.2V, 1.5V et 2V. Les 12 dispositifs ont donc subi 400 cycles d'écriture effacement pour chaque combinaison Icc/V_{top}. Pour toutes les mesures la tension de reset est laissée fixe à 2V, et le pulse est de 1µs. En figure II.35, nous avons représenté les distributions obtenues sur les 400 cycles sur les 12 dispositifs (soit 4800 points), pour le courant de compliance de 400µA (le résultat est similaire pour les autres courants de compliance).

On ne constate aucune différence significative entre ces trois distributions. Cela confirme bien que la tension de set ne joue pas un rôle prépondérant dans la dispersion des distributions de résistances, vu le rôle joué par le transistor, qui va concentrer la majorité de la tension électrique. Nous pouvons donc passer à l'étude de l'impact du temps de pulse t_{SET}, ainsi que du courant de compliance Icc.

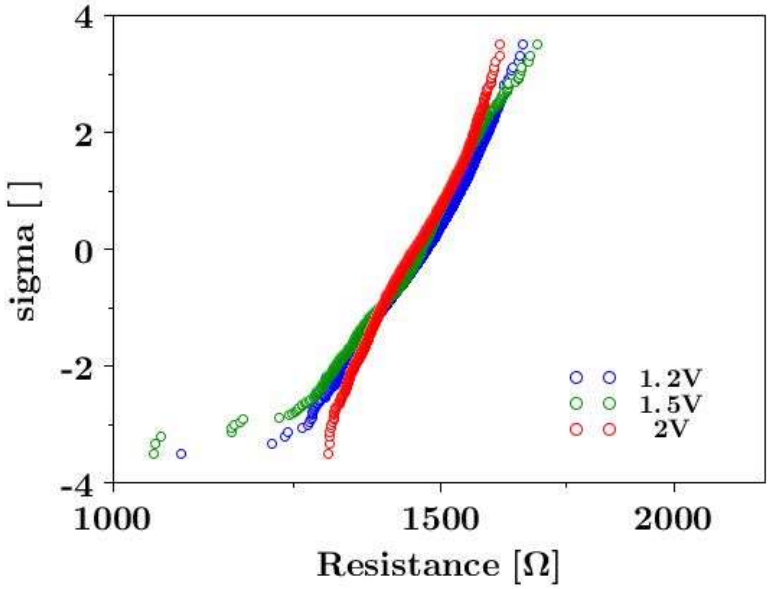


Figure II.35 Distributions LRS obtenues pour 400 cycles sur 12 cellules, pour des tensions de set de 1.2V, 1.5V et 2V.

5.3. Résultats en fonction de t_{SET} et Icc sur les distributions LRS

5.3.1/ Etude sur les dispositifs à transistors T3

Pour étudier l'influence de ces deux facteurs, nous avons décidé de réaliser des mesures similaires (400 cycles d'écriture-effacement) sur 20 devices, toujours de type T3.

Nous avons choisi d'étudier trois courants de compliance (200, 300 et 400µA) et six temps de set (20ns, 50ns, 200ns, 800ns, 2µs et 10µs). Entre chaque cycle, une opération de lecture de courant à 0.1V est réalisée. On obtient donc, pour chaque configuration Icc/t_{SET} et pour chaque cellule, un fichier de 400 lectures de courant.

Ensuite, nous rassemblons toutes les mesures correspondant à chaque configuration Icc/t_{SET}, afin de cumuler les variabilités cycle-à-cycle et device-à-device. En Figure II.36, nous avons représenté les distributions obtenues pour les trois courants de compliance et pour trois temps de pulses (20ns, 800ns et 10µs).

On constate que les distributions sont à peu près gaussiennes (même si on peut remarquer des queues de distributions pour les courbes associées à un courant de compliance de 200µA), puisqu'elles sont à peu près linéaires. L'impact du courant de compliance est très net : augmenter le courant de compliance augmente très clairement le courant de lecture (c'est-à-dire diminue la résistance de l'état LRS). Cependant on constate également que la durée du

pulse joue un rôle non négligeable puisque le fait d'augmenter la durée de pulse améliore aussi l'état LRS obtenu (mais avec un impact plus faible).

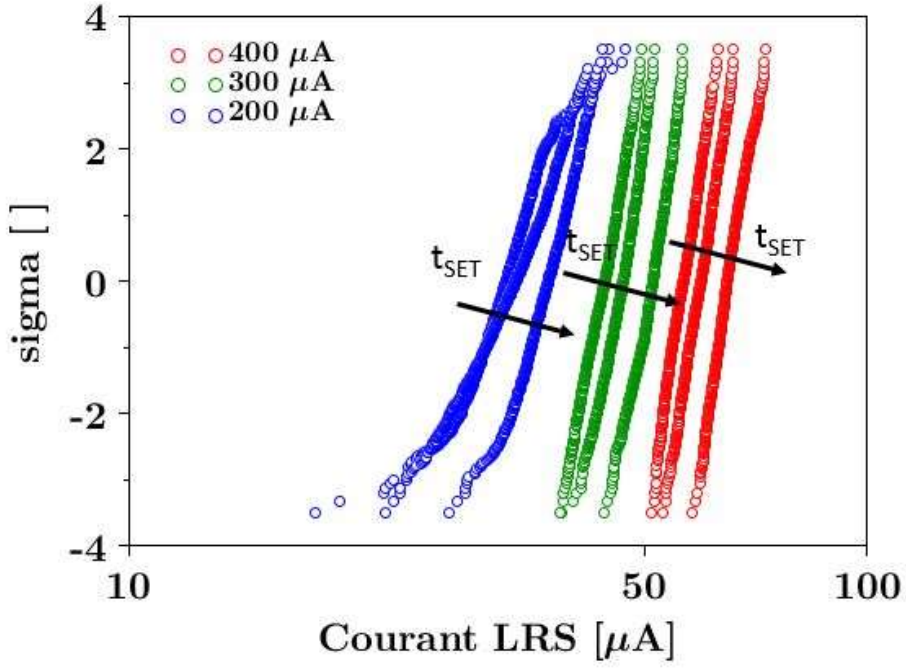


Figure II.36 : Distributions des courants de lecture LRS pour 3 courants de compliance, pour des temps de pulse de 20ns, 800ns et 10µs.

Les courbes sont à peu près parallèles, ce qui signifie que l'écart type ne semble pas varier, ni avec le courant de compliance, ni avec la durée du pulse. En revanche on peut tracer la courbe de l'évolution de la valeur moyenne de ces distributions (cf. Figure II.37).

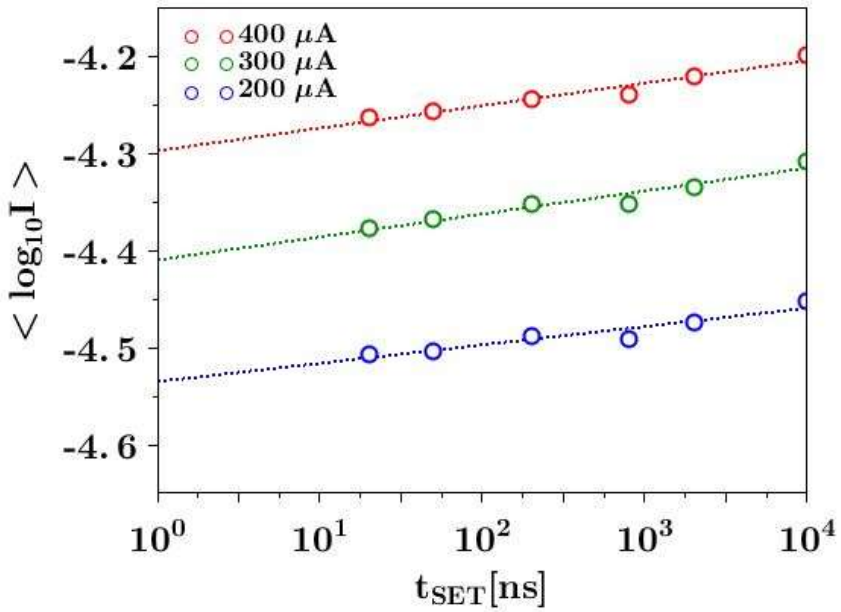


Figure II.37 : Evolution de la valeur moyenne du logarithme décimal du courant de lecture LRS pour différentes durée de pulse et différents courants de compliance.

Les trois courbes obtenues (pour chaque valeur de courant de compliance) peuvent être modélisées par des droites. Ces trois droites sont à peu près parallèles : l'influence de la durée de pulse sur la valeur du courant de lecture ne varie donc pas avec le courant de compliance.

Cependant les courants associés aux transistors T3 sont relativement élevés. Afin d'approfondir cette analyse nous avons décidé de réaliser les mêmes mesures, mais sur les dispositifs munis d'un transistor T1 (W=0.35µm).

5.3.2/ Etude sur les dispositifs à transistors T1

Le protocole de mesure utilisé ici est exactement le même que pour les dispositifs T3. Nous avons cependant décidé d'élargir les gammes de courants et de temps étudiés. Ainsi nous allons ici explorer six courants de compliance différents (20, 30, 40, 50, 60 et 80µA) et huit temps de pulse (20, 50, 200, 800ns, 2, 10, 100 et 1000µs). Là encore, 20 dispositifs ont été testés.

Nous avons donc tracé la même courbe que celle représentée en Fig. II.37 (cf. Fig. II.38).

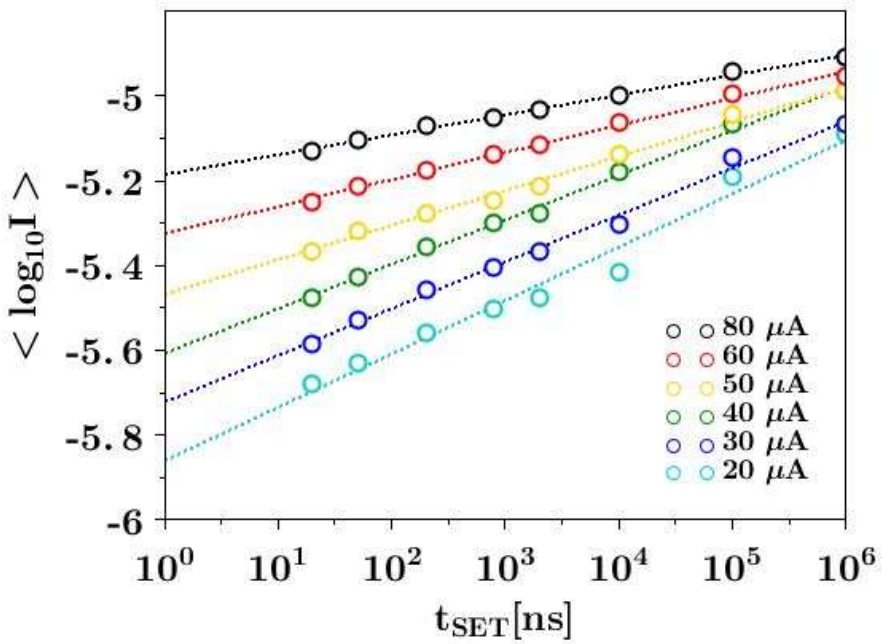


Figure II.38 : Evolution de la valeur moyenne du logarithme décimal du courant de lecture LRS pour différentes durée de pulse et différents courants de compliance.

Là encore, pour les six courants de compliance étudiés, l'influence de t_{SET} peut être modélisée par une droite (en échelle logarithmique). On remarque cependant que plus le courant de compliance augmente, moins la pente de cette droite est élevée. On en déduit qu'un fort courant de compliance réduit l'impact du temps de set. En revanche, pour un courant de 20µA

par exemple, le passage d'un temps de pulse de 20ns à 1ms change de façon importante la valeur moyenne du courant obtenue.

5.3.3/ Mise en commun des résultats et modélisation

En figure II.39 nous avons tracé les deux courbes précédentes sur un même graphique.

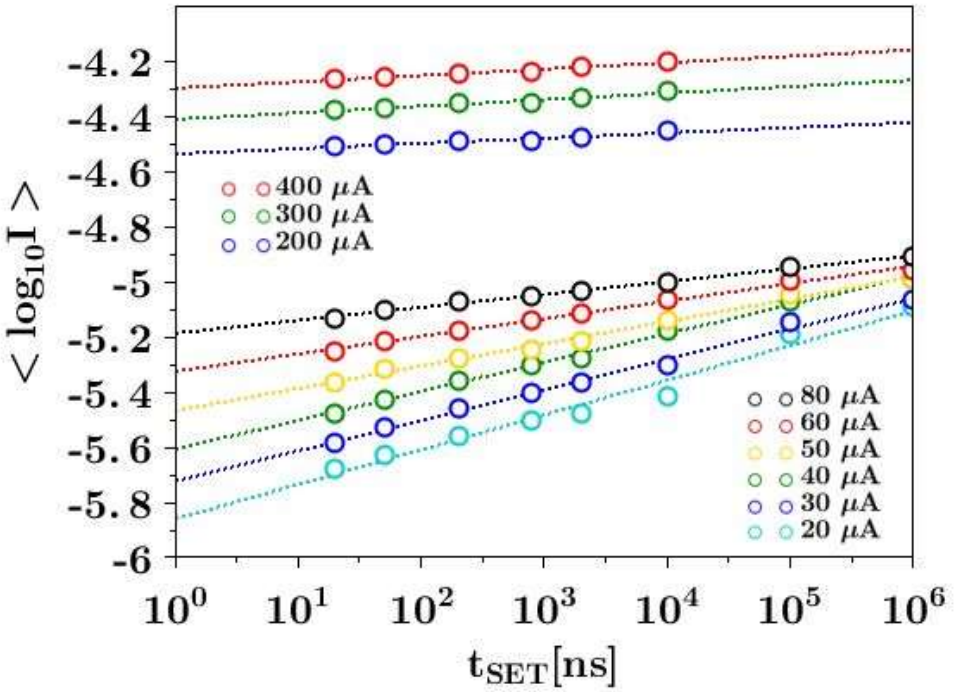


Figure II.39 : Mise en commun des courbes réalisées à faibles et forts courants.

Une tendance assez nette se dégage de ces courbes et peut être modélisée par :

$$< log_{10}(I) > = a(Icc).log_{10}(t_{SET}) + b(Icc)$$

Plus le courant augmente, plus l'ordonnée à l'origine augmente (ce qui signifie simplement que plus le courant augmente, plus les états LRS obtenus sont bons). La pente des fits, quant à elle, diminue avec le courant, ce qui laisse supposer que l'influence du temps de pulse diminue lorsque le courant est important. Nous avons alors tracé les évolutions de cette pente (notée simplement a) et l'ordonnée à l'origine (notée b) en fonction du courant de compliance utilisé (cf. Fig. II.40 (a) et (b)). Comme on peut le voir, l'évolution de la pente peut être fittée par une exponentielle décroissante. Cette pente quantifie l'influence qu'a le temps de pulse sur l'état LRS obtenu. Cette influence diminue rapidement avec l'augmentation du courant, jusqu'à environ 100µA, où elle se stabilise. L'ordonnée à l'origine b a la forme d'une courbe de saturation. Mathématiquement, b représente la valeur théorique de $<\log_{10}I>$ pour un temps de pulse de 1ns. Alors que celle-ci est très sensible à des variations de courant lorsque celui-ci est inférieur à 200µA, on observe une certaine saturation qui laisse supposer que, passé

une certaines valeur de courant de compliance, continuer à augmenter ce courant, n'améliore plus significativement les états LRS obtenus.

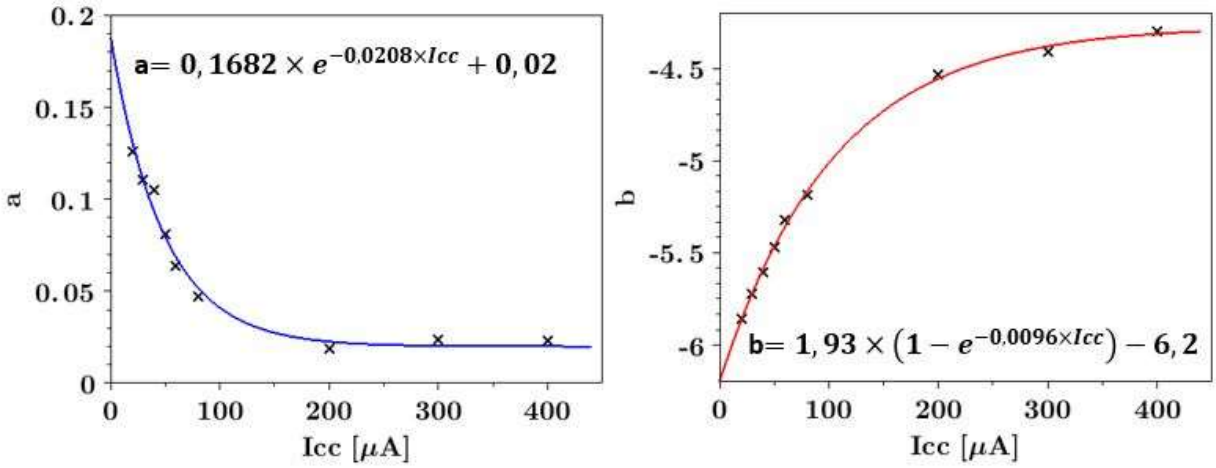


Figure II.40 : Valeur de la pente et de l'ordonnée à l'origine des courbes de fit de la valeur moyenne du courant de lecture LRS.

Nous avons donc compris l'évolution de la valeur moyenne des distributions des courants de lecture LRS. Nous pouvons également regarder comment évolue l'écart type σ de ces distributions. Au contraire de la valeur moyenne, où, pour chaque courant de compliance, nous avons choisi de tracer les courbes en fonction du temps de pulse, l'évolution de σ était beaucoup plus flagrante lorsque nous la traçons en fonction du courant de compliance (cf. Fig. II.41).

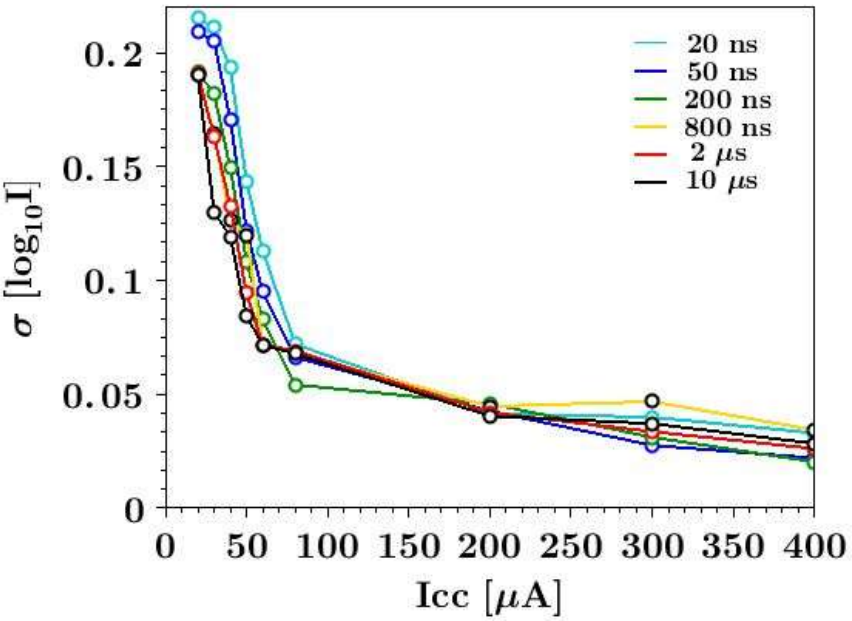


Figure II.41 : Evolution de l'écart type des distributions de courants de lecture LRS en fonction du courant de compliance utilisé, pour différentes longueurs de pulse.

En Figure II.42, nous avons modélisé ces courbes par des exponentielles décroissantes. Le paramètre B est représenté en fonction de t_{SET} en inset. Le paramètre A est presque constant à 0.3 pour tous les t_{SET}.

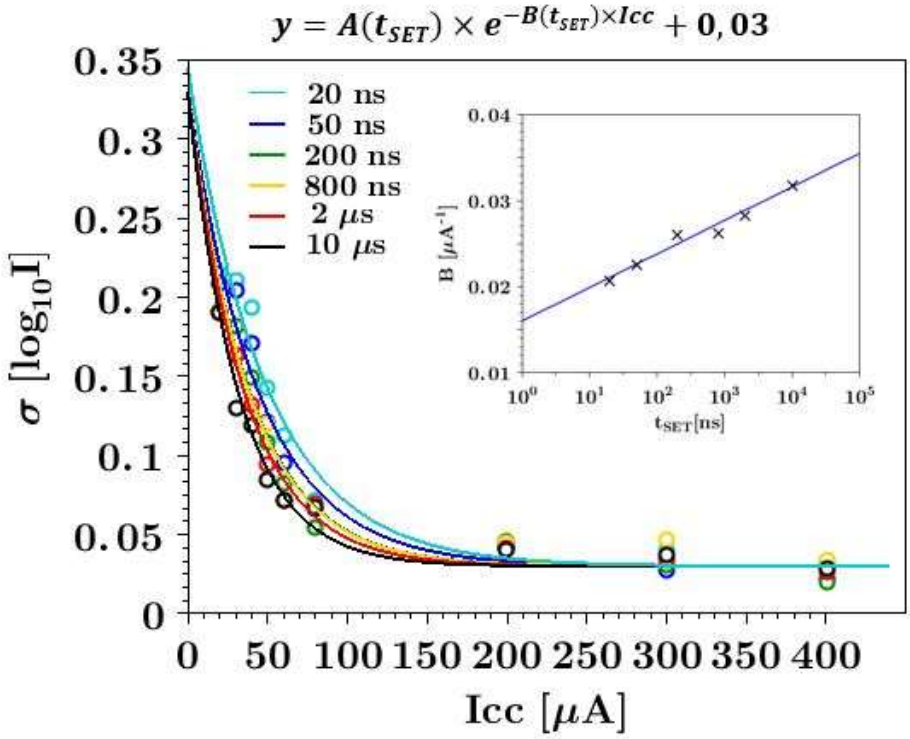


Figure II.42 : Fit des courbes de σ en fonction d'Icc pour différents t_{SET}. On remarque que l'évolution de B évolue linéairement avec le logarithme décimal du temps de pulse.

On constate que l'influence du temps de pulse est moins importante que celle du courant de compliance. En effet les six courbes sont relativement similaires. On remarque en revanche que l'influence du courant de compliance est très nette. A faible courant, augmenter le courant permet de réduire de façon très importante l'écart type, peu importe le temps de pulse utilisé. Au-delà de 100µA, l'écart se stabilise.

Grâce à ces mesures, nous avons déduit des formules empiriques permettant de modéliser les distributions de courants de lecture LRS, pour n'importe quel temps de pulse et courants de compliance.

Plusieurs informations peuvent être déduites de ces mesures :

- Au-delà d'une certaine valeur de courant, à la fois l'écart-type des distributions ne s'améliore plus beaucoup, et les facteurs a et b modélisant l'évolution de la valeur moyenne des distributions avec le temps de pulse se stabilisent. On en déduit que dépasser les 100µA, voir 200µA ne semble pas particulièrement utile pour améliorer les distributions des états LRS.

- Que ce soit pour l'écart type ou la valeur moyenne des distributions, l'influence du courant de compliance est nettement plus importante que celle du temps de pulse. Il semble donc tout à fait possible d'utiliser des temps de pulse très faibles, quitte à compenser avec une légère augmentation du courant de compliance. En effet, d'un point de vue énergétique, baisser de plusieurs ordres de grandeur le temps de pulse compense très largement une légère augmentation du courant de compliance utilisé. Cela est illustré en Fig. II.43. Nous avons comparé les distributions totales (400 cycles sur 20 cellules) pour des sets effectués avec des pulses de 800ns à une compliance de 20µA contre des pulses de 20ns à une compliance de 80µA. On peut noter que le fait de diminuer le temps de pulse d'un facteur 40 est très largement compensé par le fait d'augmenter le courant de compliance d'un facteur 4. On obtient ainsi une distribution bien meilleure en augmentant le courant de compliance et en travaillant sur des pulses beaucoup plus courts. D'un point de vue énergétique, cette opération est totalement bénéfique. En effet l'énergie utilisée pendant le set s'exprime ainsi :

$$E_{set} = \int_0^{t_{SET}} I_{set} V_{set} dt$$

Or, grâce à l'étude en dynamique, on sait que le switching intervient très tôt dans l'opération. On peut donc approximer l'énergie ainsi : $E_{set} = I_{set} V_{set} t_{SET}$. Ainsi, en plus d'améliorer très nettement les distributions obtenues, on diminue de cette façon l'énergie consommée d'un facteur 10.

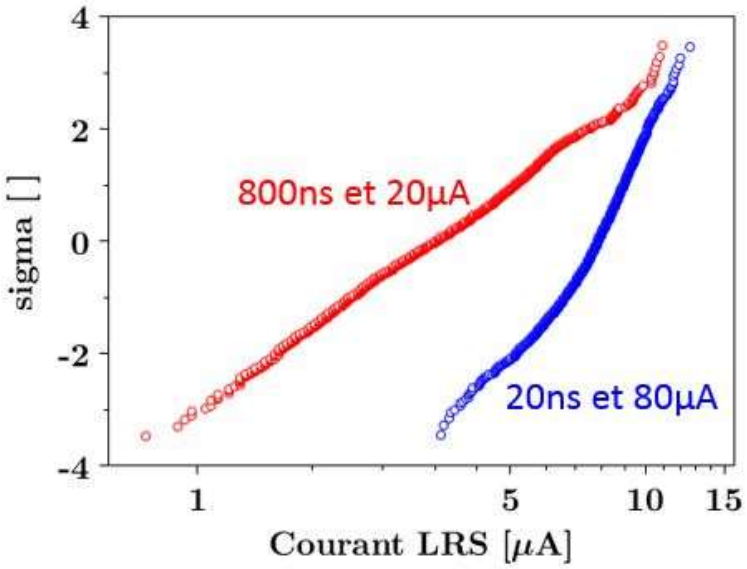


Figure II.43 : Comparaison entre une distribution obtenue avec des pulses de 800ns à 20µA et une autre obtenue avec des pulses de 20ns à 80µA.

5.4. Equivalent pour le reset : influence de t_{RESET} et de Vr

A la différence de l'opération de set, qui est limitée par le courant de compliance, on ne cherche pas à limiter le courant lors du reset. C'est donc la tension de reset qui fixe l'intensité de l'opération, et l'état HRS obtenu. Pour ces manipulations nous avons donc utilisé le même protocole expérimental (même banc de test, 400 cycles sur 20 cellules). Nous avons toujours fait varier le temps de pulse (20, 50, 200, 800 ns, 2, 10 et 100µs). Nous avons simplement fait varier la tension de reset à la place du courant de compliance (nous avons testé des tensions de reset de 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 2 et 2.2V). Les opérations de set ont été réalisées à un courant de compliance de 50µA, sur des pulses de 1µs. Nous avons ensuite analysé les distributions de courant de lecture des états HRS obtenus.

Comme pour les distributions LRS, nous avons tracé et modélisé l'évolution de la valeur moyenne du logarithme décimal du courant de lecture en fonction du temps de pulse, pour chaque tension de reset étudiée. Ces courbes sont représentées en Figure II.44.

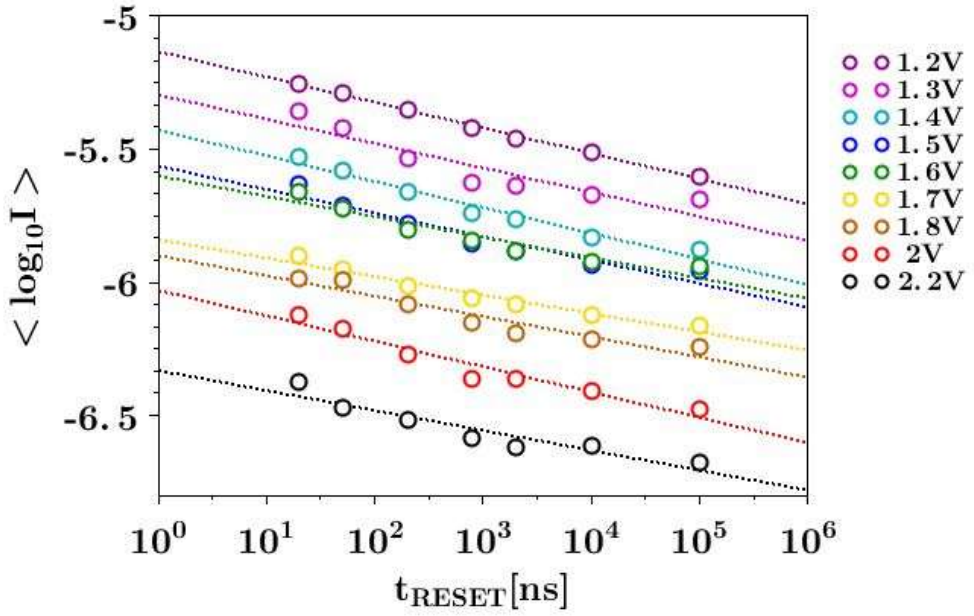


Figure II.44 : Evolution de la valeur moyenne du logarithme décimal du courant de lecture HRS pour différentes durées de pulse et différentes tensions de reset.

On constate que l'influence du temps de pulse de reset est vraiment similaire à celle du temps de pulse de set, puisque l'on obtient également des droites que nous avons modélisées par l'équation : $< log_{10}(I) > = a(Vr).log_{10}(t_{RESET}) + b(Vr)$. Les pentes a et ordonnées à l'origine b de ces droite sont tracées en fonction de la tension de reset en Figure II.45. Pour rappel, comme nous étudions des états HRS, nous visons des courants de lecture faibles. De même que pour le set, on remarque que l'influence du temps de pulse est du second ordre,

puisque c'est surtout l'augmentation de la tension de reset qui permet d'améliorer les courants obtenus.

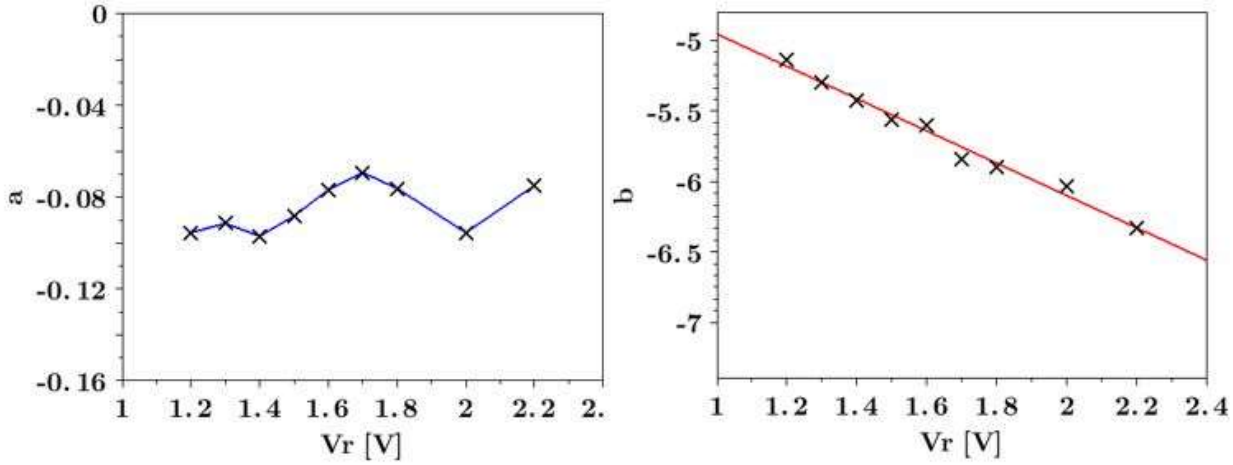


Figure II.45 : Valeur de la pente a et de l'ordonnée à l'origine b des courbes de fit de la valeur moyenne du courant de lecture HRS.

La pente des courbes de la Figure II.44 ne semble pas dépendre de la tension de reset (cf. Figure II.45, et le fait que les courbes sont à peu près parallèles). Augmenter la tension de reset impacte donc principalement l'ordonnée à l'origine b de ces droites. Nous avons pu modéliser l'influence de Vr sur b par une droite. b étant de la dimension du logarithme du courant, on en déduit un impact en exponentiel de la tension de reset sur la résistance HRS obtenue.

Comme pour le set, nous avons également tracé l'évolution de l'écart type des distributions en fonction de la tension de reset et du temps de pulse, en Figure II.46.

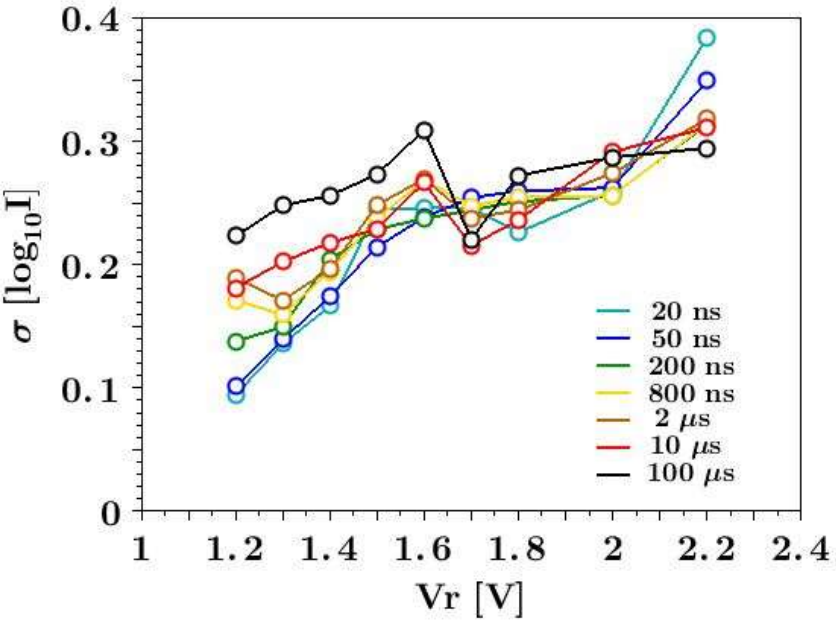


Figure II.46 : Evolution de l'écart type des distributions de courants de lecture HRS en fonction de la tension de reset utilisée, pour différentes longueurs de pulse.

Même si aucune réelle tendance ne ressort clairement, il apparaît que l'augmentation de la tension de reset augmente l'écart type des distributions. Cependant, cela est très nettement compensé par l'évolution de la valeur moyenne (de plus de 1, en échelle log, alors que l'augmentation de l'écart type n'est que de 0.3 au maximum).

Le temps de pulse, en revanche ne semble pas avoir d'impact net sur l'écart type des distributions.

Au final, la conclusion est la même que pour le set : il est plus intéressant, autant d'un point de vue énergétique qu'au niveau de la qualité des distributions de diminuer la longueur de pulse, quitte à moins chercher à diminuer les courants ou tensions employées. En Figure II.47, nous avons comparé les distributions obtenues pour des pulses de reset relativement courts, de 800ns avec une tension de 1.8V, à celles obtenues avec des pulses beaucoup plus longs de 100µs, avec une tension de 1.7V. On remarque que les deux distributions sont quasiment similaires. Ainsi, en diminuant drastiquement la longueur des pulses (d'un facteur 125), et en n'augmentant la tension de seulement 0.1V, on conserve exactement la même variabilité.

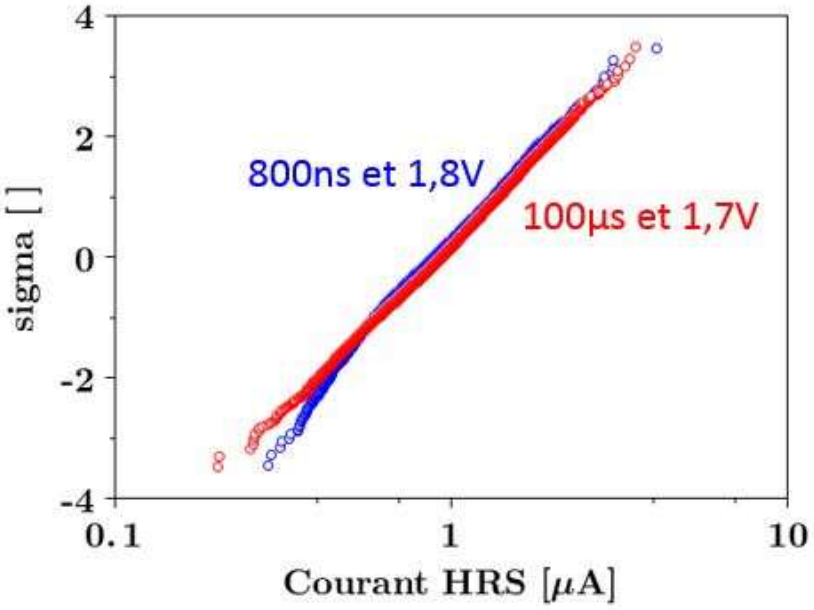


Figure II.47 : Comparaison entre une distribution obtenue avec des pulses de 800ns à 1.8V et une autre obtenue avec des pulses de 100µs à 1.7V.

6. Conclusion

Dans cette partie, exclusivement expérimentale, nous avons démontré tout d'abord la faisabilité d'opérer sur des temps très courts, jusqu'à des pulses de quelques nanosecondes.

Grâce à une architecture de cellules mémoires 1T1R particulière, dotée d'un point d'accès entre le transistor et la cellule mémoire, nous avons été capables de suivre de façon dynamique le processus de switching, jusqu'à ces temps très courts. Cela nous a permis de tracer la courbe d'évolution des tension de set et de reset en fonction de la vitesse de la rampe utilisée, ce qui n'avait jamais été présenté dans la littérature, jusque-là, sur des structures 1T1R.

Ensuite, nous avons démontré l'intérêt d'avoir des dispositifs capables d'opérer sur des temps très courts, puisque la réduction des temps de pulses n'affecte que peu la variabilité des dispositifs. En effet, la réduction de plusieurs ordres de grandeurs de la taille des pulse est très facilement compensée par une légère augmentation du courant de compliance utilisé (pour l'opération de set) et de la tension de reset employée (pour l'opération de reset).

Cette étude nous a permis de réaliser l'un des objectifs de la thèse, à savoir la démonstration de la faisabilité d'un fonctionnement sur des temps courts.

Dans un deuxième temps, cet accès à la dynamique du switching nous offre la possibilité de nous interroger sur les mécanismes physiques associés au phénomène de commutation à la base du fonctionnement des OxRAM. En effet, en s'appuyant sur ces résultats, nous pouvons maintenant passer à la partie consacrée à la modélisation physique des phases d'écriture et d'effacement.

7. Références du chapitre II

[1] http://www.tek.com/sites/tek.com/files/media/media/resources/4200A-SCS_Datasheet_1KW-60780-2.pdf

[2] https://smt.at/wp-content/uploads/smt-datenblatt-keithley-4225-englisch.pdf

[3] http://www.pd.infn.it/elettronica/Strumenti/HP8110A.pdf

[4] J. Hutchby and M. Garner, "Assessment of the potential & maturity of selected emerging research memory technologies," ITRS Workshop & ERD/ERM Working Group Meeting, 2010

[5] K. Kinoshita et al., "Reduction in the reset current in a resistive random access memory consisting of NiOx brought about by reducing a parasitic capacitance" APPLIED PHYSICS LETTERS 93, 2008

[6] H.-S.P. Wong, H.-Y. Lee, S. Yu et al., "Metal–Oxide RRAM," Proceedings IEEE, pp. 1951-1970, 2012

[7] S. Larentis et al., "Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling" IEEE Transactions on Electron Devices, vol. 59, no. 9, September 2012

[8] A. Padovani et al., “Microscopic modeling of HfOx RRAM operations: from forming to switching”, IEEE Transactions on Electron Devices, vol. 62, no. 6, June 2015

[9] L. Goux et al., " Role of the Ta scavenger electrode in the excellent switching control and reliability of a scalable low-current operated TiN\Ta2O5\Ta RRAM device”, Very Large Scale Integr. (VLSI), 2014

[10] S-S. Sheu et al., “A 5ns fast write multi-level non-volatile 1 K bits RRAM memory with advance write scheme”, Very Large Scale Integr. (VLSI), 2009

[11] K. Tsunoda et al., “Low power and high speed switching of Ti-doped NiO ReRAM under the unipolar voltage source of less than 3 V”, IEEE International Electron Devices Meeting (IEDM) 2007, TECHNICAL DIGEST, 2007

[12] Y.S. Chen et al., “Robust high-resistance state and improved endurance of HfOX resistive memory by suppression of current overshoot”, IEEE Electron Device Letter, vol.32, no.11, November 2011

[13] B. Govoreanu et al., “Performance and Reliability of Ultra-Thin HfO2-Based RRAM (UTO-RRAM)”, 5TH IEEE International Memory Workshop (IMW), 2013

[14] http://www.tek.com/sites/tek.com/files/media/media/resources/TPP1000-TPP0500B-TPP0502-TPP0250-Probe-Datasheet-51W261517_1.pdf

[15] D. Ielmini, C. Cagli, and F. Nardi, “Resistance transition in metal oxides induced by electronic threshold switching” Applied Physics Letters 94, 063511, 2009

[16] H. B. Lv et al., “Resistive memory switching of CuxO films for a nonvolatile memory application”, IEEE Electron Device Letter, vol. 29, no. 4, pp. 309–311, April 2008

[17] M. G. Cao et al., “Nonlinear dependence of set time on pulse voltage caused by thermal accelerated breakdown in the Ti/HfO2/Pt resistive switching devices”, Applied Physics Letters 101, 203502, 2012

[18] C. Cagli, F. Nardi, and D. Ielmini, “Modeling of Set/Reset operations in NiO-based resistive-switching memory devices” IEEE Transactions on electron devices Volume: 56 Issue: 8 Pages: 1712-1720, August 2009

[19] X. Guan et al., “On the Switching Parameter Variation of Metal-Oxide RRAM _ Part I: Physical Modeling and Simulation Methodology” IEEE Transactions on electron devices Volume: 59 Issue: 4 Pages: 1172-1182, April 2012

[20] C. Nguyen, C. Cagli, E. Vianello, A. Persico, G. Molas, G. Reimbold, Q. Rafhay and G. Ghibaudo, “Advanced 1T1R test vehicle for RRAM nanosecond-range switching-time resolution and reliability assessment”, IEEE International Integrated Reliability Workshop, 2015

[21] J.W. McPherson, J. Y. Kim, A. Shanware, H. Mogul, “Thermochemical description of dielectric breakdown in high dielectric constant materials”, Appl. Phys. Lett., Vol 82(13), pp. 2121-2123, 2003

[22] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy and D. Ielmini, “Statistical Fluctuations in HfOx resistive-switching memory: part I – Set/Reset variability”, IEEE Transactions on Electron Devices, vol. 31, no. 8, august 2014

[23] B. Govoreanu et al., “Complementary role of field and temperature in triggering ON/OFF switching mechanisms in Hf/HfO2 Resistive RAM cells”, IEEE Transactions on Electron Devices, Vol. 60, No. 8, August 2013

[24] S. Ambrogio, S. Balatti, DC. Gilmer and D. Ielmini, “Analytical modeling of Oxide-based bipolar Resistive Memories and complementary resistive switches IEEE Transactions on Electron Devices, Vol. 61, No. 7, July 2014

[25] D. Berco and T-Y. Tseng, “A comprehensive study of bipolar operation in resistive switching memory devices”, Journal of Computation al Electronics, Volume 15, Issue 2, pp 577–585, June 2016

[26] D. Ielmini, “Modeling the universal Set/Reset characteristics of filament growth by field- and temperature- driven filament growth,” IEEE Transactions on Electron Devices, vol. 58, no. 12, pp. 4309–4317, December 2011

[27] D. Garbin et al., “Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network”, Eurosoi ULIS pp.125-128, 26-28 Jan. 2015

[28] A. Fantini et al., “Intrinsic switching variability in HfO_2 RRAM”, 5th International Memory Workshop (IMW), 2013

Chapitre III : Etude de la dynamique de switching via un modèle semi-analytique

1. Introduction

Ce troisième chapitre du rapport de thèse a pour objet la modélisation physique des opérations d'écriture et d'effacement sur les cellules mémoires OxRAM. En particulier nous recherchons à obtenir un modèle physique capable de reproduire les courbes de switching en quasi-statique et également en dynamique sur les temps très courts que nous avons réalisées dans le chapitre précédent. D'un autre côté, nous désirons un modèle restant suffisamment simple afin de pouvoir calculer rapidement les courbes de commutation. En effet nous voulons être capables de reproduire des courbes de statistique sur plusieurs dispositifs ou plusieurs cycles, afin d'essayer de prévoir la fiabilité et la variabilité de cellules mémoires.

On trouve dans la littérature un grand nombre de modèles différents. On peut classer ces modèles en deux grandes catégories. La première est la catégorie dite « memristor-like » qui consiste à modéliser le comportement de la mémoire par une variable d'état, sans chercher à rentrer dans les détails microscopiques du fonctionnement de la mémoire [1, 2, 3]. La seconde catégorie est celle des modèles physiques qui prennent en compte les phénomènes de génération, dissolution et migration d'un filament conducteur, à l'échelle atomique [4, 5, 6]. Nous cherchons donc un modèle situé à la frontière entre ces deux catégories, associant la vitesse d'exécution de la première catégorie et l'accès aux paramètres physiques de la seconde catégorie.

Nous commencerons cette partie par décrire le contexte du modèle. Nous établirons les principales hypothèses utilisées pour la suite, ainsi que les grandeurs physiques sur lesquelles le modèle repose.

Ensuite nous nous poserons la question du type de conduction mise en jeu pour les différents états possibles de la cellule. Pour cela, nous réaliserons une étude de l'influence de la température sur la résistance des états LRS et HRS. En effet, l'étude de l'activation en température permet de discriminer quels sont les modes de conduction possibles.

Ensuite nous détaillerons comment nous passerons d'un paramètre purement physique (le nombre de lacunes d'oxygène présentes) à un paramètre électrique, directement mesurable (la résistance de la cellule).

Nous tenterons ensuite de modéliser les étapes de switching en elles-mêmes et les phénomènes physiques associés (la génération de lacune d'oxygène, pour l'écriture, et leur recombinaison/migration pour l'effacement).

Tout au long de la mise en place du modèle, nous confronterons nos hypothèses à l'expérimentation, ce qui nous amènera à considérer la nécessité d'implémenter la notion de compliance dans notre modèle, et à s'intéresser de près au calcul de la température au sein du filament conducteur.

Une fois le modèle mis au point, nous l'appliquerons pour tenter de reproduire les courbes obtenues dans la partie précédente, ce qui nous permettra d'étalonner notre modèle, c'est-à-dire de fixer un certain nombre de paramètres, qui ne peuvent jusque-là être qu'arbitraires. Enfin nous conclurons à la fois sur la gamme d'application du modèle et sur son efficacité à reproduire la dynamique de switching, en particulier sur les temps très courts.

2. Mise au point d’un modèle de switching

Dans la mesure où nous recherchons un modèle principalement analytique (c'est-à-dire qui peut être décrit par un certain ensemble d'équations physiques et mathématiques), nous avons choisi de coder ce modèle sur Scilab [7]. Nous n'aurons pas besoin de définir un maillage, qui nécessiterait des ressources informatiques plus importantes, et qui nous ferait perdre en vitesse d'exécution.

2.1. Contexte du modèle

Le but de ce modèle est de représenter le fonctionnement d'une cellule OxRAM (c'est-à-dire la structure MIM capable de commuter entre un état hautement résistif HRS et un état faiblement résistif LRS). Ainsi, "l'environnement" du modèle sera tout simplement une

structure MIM : une couche d'oxyde capable de switcher prise en sandwich entre deux électrodes métalliques (cf. Fig. III.1).

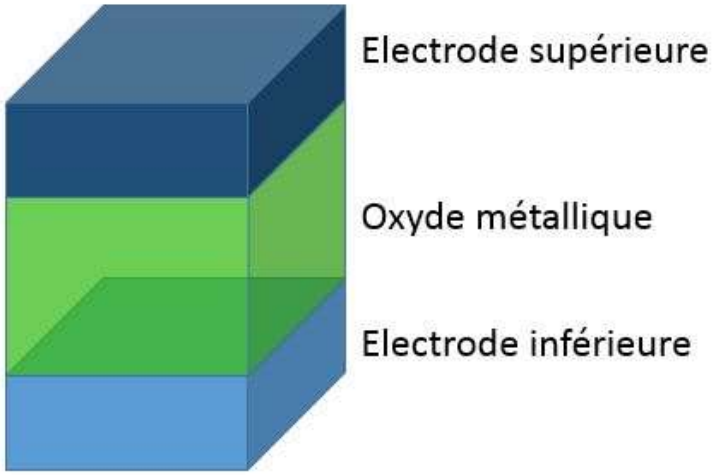


Figure III.1 : Environnement du modèle : structure MIM classique

Nous supposerons, comme c'est le cas dans la plupart des papiers traitant de la modélisation des OxRAM, que la conduction repose sur la présence d'un filament conducteur constitué de lacunes d'oxygène [8].

Le but de ce modèle est de décrire les phases de set et de reset. Ainsi, nous considèrerons que le filament conducteur est déjà présent. En revanche, en fonction de l'état initial HRS ou LRS, celui-ci peut être partiellement rompu ou non. L'opération de set va consister à générer, au sein de l'oxyde, des lacunes d'oxygène qui vont contribuer à reconstituer le filament, ou le renforcer, et à augmenter la conductance du dispositif. Au contraire, le reset consiste à supprimer un certain nombre de ces lacunes d'oxygène, pour augmenter la résistance du dispositif. On note alors g, la longueur sur laquelle le filament est rompu. La région de longueur g constitue la "zone de travail" du modèle : c'est dans cette zone que l'on va générer ou enlever des lacunes d'oxygène. Ainsi, on va supposer que le reste du filament va rester stable tout au long de l'opération. Cette hypothèse est relativement réaliste puisqu'elle modélise le caractère irréversible de l'opération de forming : une fois formée, une cellule ne retrouve jamais sa résistance de pré-forming.

On notera r le rayon du filament, et L l'épaisseur d'oxyde qui sépare les deux électrodes métalliques. Le rayon r du filament sera également supposé constant tout au long de l'opération à réaliser. La Figure III.2 schématise le principe de base de switching utilisé pour ce modèle, pour les étapes d'écriture et d'effacement.

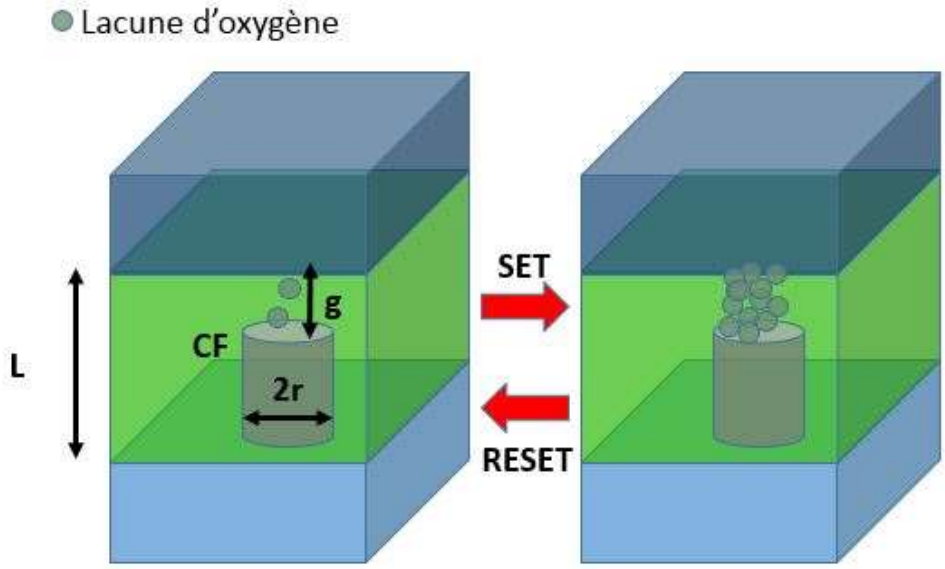


Figure III.2 : Schématisation des opérations de set (création de lacunes) et de reset (suppression de lacunes), et l'intérêt de la zone de gap g, où sont créées/supprimées les lacunes.

L est l'épaisseur d'oxyde, donc sa valeur est connue (en général, soit 5nm soit 10 nm, en fonction des échantillons). En l'absence de données expérimentales nous ne pouvons en revanche pour l'instant que faire des suppositions sur la valeur de g (qui peut aller de 0 à L) de l'ordre de quelques nanomètres. De même nous supposerons pour l'instant un rayon r de quelques nanomètres également.

Avant de traiter de la génération et recombinaison de lacunes, nous allons tout d'abord nous interroger sur la nature de la conduction.

2.2. Etude de l'activation de la conduction en température

Comme nous l'avons vu dans la partie dédiée à la modélisation, du premier chapitre, plusieurs modes de conduction permettent de modéliser la conduction dans les OxRAM. Les principales pistes pour le régime HRS sont la conduction tunnel assistée par pièges [9, 10, 11] et la conduction Poole-Frenkel [12, 13, 14].

Cependant les équations régissant la conduction tunnel assistée par piège (TAT) sont assez lourdes [11, 15, 16], et prennent en compte le calcul de chemins de percolation, ce qui nécessite d'associer à chaque piège une position dans la zone de travail. Or pour garder un modèle simple et rapide nous préférons nous limiter à calculer la densité de lacunes présentes dans le gap. C'est pourquoi nous nous sommes plutôt pencher vers la piste Poole-Frenkel.

Nous nous sommes beaucoup appuyés sur les travaux de l'équipe de Daniele Ielmini, du groupe de Politecnico di Milano. Dans [17, 18, 19], il utilise un modèle analytique Poole-Frenkel pour modéliser le courant dans des dispositifs PCRAM à base de chalcogénures (cf. Fig. III.3).

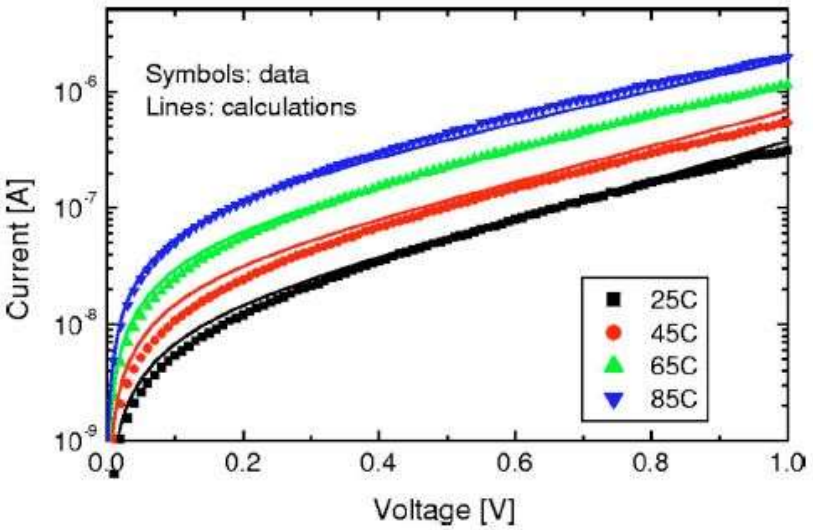


Figure III.3 : Courants expérimentaux et simulés via Poole-Frenkel sur des dispositifs de chalcogénures amorphes, à différentes températures [17].

Lorsqu'on essaye de prouver que la conduction obéit à un régime Poole-Frenkel, on réalise, comme sur la Figure III.3 des mesures en température. En effet, le mode de conduction Poole-Frenkel implique une résistance de la forme $\boldsymbol{R = A.e^{\frac{E_{AC}}{kT}}}$ (1), exponentiellement activée en température. Ainsi, nous avons sélectionné une cinquantaine de cellules, que nous avons formées puis effacées ou écrites à différentes résistance. Ensuite, nous avons mesuré la résistance de chacune de ces cellules, sous une tension de 0.1V, pour différentes températures (25, 50, 75, 100, 125 et 150°C). Pour chaque cellule on peut ensuite tracer l'évolution de cette résistance, en échelle logarithmique. On réalise alors un fit linéaire de cette évolution. On peut alors, grâce à l'équation (1) trouver l'énergie d'activation de conduction correspondante via la pente de ce fit. En Figure III.4, nous avons représenté l'évolution en température de la résistance d'un dispositif, de résistance de 140kΩ à température ambiante et le fit correspondant. L'énergie d'activation associée est de 0.0614eV.

Nous pouvons alors tracer, pour tous les dispositifs étudiés, l'évolution de cette énergie d'activation de conduction en fonction de la résistance du dispositif (ici nous définissons cette résistance comme celle lue à 0.1V à température ambiante). A noter également que nous avons choisi de ne pas former certaines cellules pour étudier la conduction lors de l'état pristine. Cette courbe est représentée en Figure III.5.

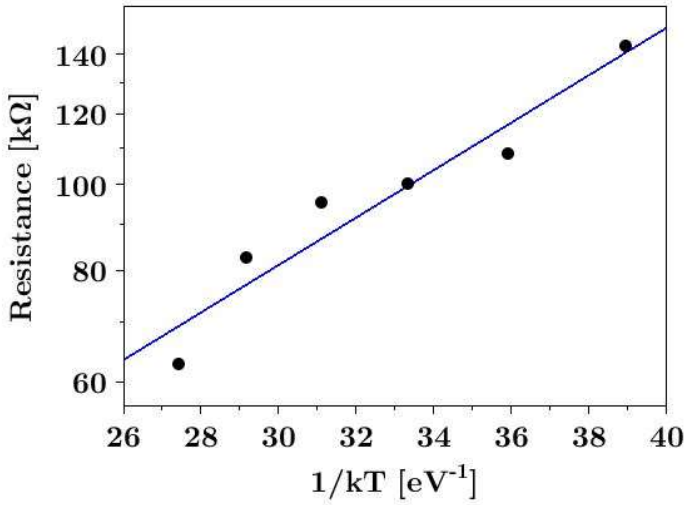


Figure III.4 : Evolution de la résistance d'un dispositif OxRAM en fonction de la température. En bleu, nous avons tracé le fit de cette évolution.

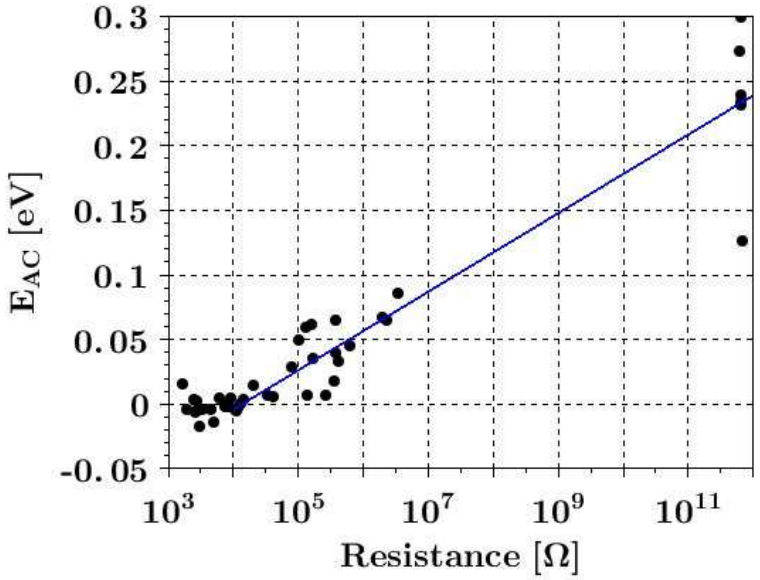


Figure III.5 : Evolution de l'énergie d'activation de conduction en fonction la résistance (à température ambiante). En bleu, nous avons modélisé cette évolution, pour les résistances supérieures à 10kΩ par une droite.

On peut très clairement identifier deux régimes différents : à faible et à forte résistance. A faible résistance (en dessous de 10kΩ), on n'observe aucune activation en température (énergie d'activation nulle), ce qui confirme un régime métallique ohmique, avec un niveau de Fermi dans la bande de conduction comme expliqué dans [13 et 20]. A forte résistance, on observe une augmentation linéaire de l'énergie d'activation de conduction en fonction du logarithme décimal de la résistance (comme c'est également le cas dans [13 et 20]). On peut ainsi, comme dans [13] comparer l'influence des lacunes d'oxygène (dont la concentration augmente lorsque la résistance diminue) à celle de dopants dans des semi-conducteurs. A

mesure que leur concentration augmente, on passe d'un état très peu conducteur, à très forte énergie d'activation (semi-conducteur intrinsèque ou faiblement dopé) à un état où le niveau de Fermi se rapproche de plus en plus de la bande de conduction, jusqu'à obtenir un régime métallique.

Nous avons également pu fitter l'évolution de l'énergie d'activation de conduction pour les résistances supérieures à $10^4\ \Omega$ par une droite dont l'équation est la suivante :

$$\boldsymbol{E_{AC} = 0.030375.log_{10} R - 0.1256eV} \quad (2)$$

Nous nous resservirons de cette équation par la suite, lors de la mise au point du modèle.

Cette étude nous permet de confirmer ce qu'on peut lire dans la majorité des papiers présents dans la littérature : la conduction au sein des OxRAM est bien séparée en deux régimes. Un régime LRS non activé en température et un régime HRS activé en température. Nous allons donc traiter ces deux régimes séparément.

2.3. Etude de la conduction de l'état HRS

Nous avons décidé de fitter la conduction de l'état HRS par un modèle de conduction Poole-Frenkel [21], parfaitement compatible avec les résultats obtenus dans la partie précédente. Une conduction Poole-Frenkel est un mode de conduction assistée par piège, et activée en température, donc applicable pour l'état HRS. Bien sûr, le rôle des pièges est ici joué par les lacunes d'oxygène.

Traditionnellement, un courant Poole-Frenkel est modélisé par une exponentielle de la racine carrée du champ électrique [17, 21] de la forme $\mathbf{I} \propto \boldsymbol{I_0}.\boldsymbol{e^{\beta.\sqrt{V}}}$.

Nous avons donc cherché à obtenir des courbes I-V de cellules OxRAM dans l'état HRS. Ceci est une opération assez délicate car nous voulons nous affranchir de tout phénomène de switching. Or pour chercher à démontrer une modélisation via un courant Poole-Frenkel nous avons besoin de monter la tension appliquée à une valeur supérieure à 1V, valeurs auxquelles le switching survient. Pour contourner ce problème, nous avons décidé de collecter nos mesures de courant sur le retour d'une opération de reset, comme indiqué en Figure III.6, au moment où il y a le moins de perturbation en courant.

Nous avons donc réalisé quarante mesures en quasi-statique, pour différentes valeurs de tension de reset (comme la tension de reset commande directement la résistance obtenue après opération, cela nous permet d'avoir accès à un panel assez large de résistance HRS). Nous avons ainsi obtenu des résistances HRS allant de quelques dizaines de kΩ à plusieurs MΩ.

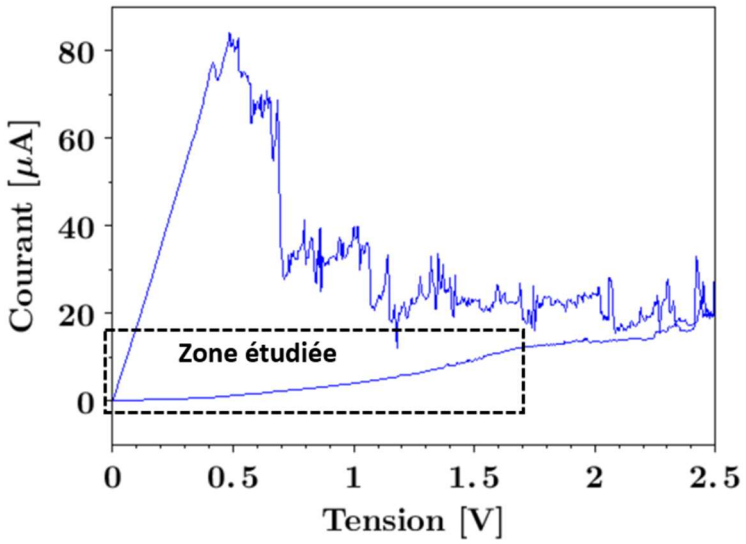


Figure III.6 : Mesure quasi-statique d'une opération de reset. Dans le rectangle en pointillé nous avons indiqué la zone dans laquelle nous collectons nos données, pour comprendre le régime de conduction de l'état HRS.

Nous avons alors pu tenter de fitter ces courbes. Nous avons tout d'abord observé qu'une simple exponentielle de la racine carrée de la tension ne permettait pas de modéliser convenablement le courant HRS. Nous avons ainsi fait la même observation que dans [17] : à faible tension (jusqu'à 150mV), le courant augmente de façon linéaire avec la tension. Ceci est visible en Figure III.7. Les deux courbes du bas, à des résistances élevées sont, en échelle log-log, parallèles à la courbe du haut, à faible résistance (et donc linéaire). Ce n'est qu'à plus forte tension que l'allure en exponentielle de la racine carrée de la tension apparaît et que les courbes du bas ne sont plus parallèles à la courbe du haut.

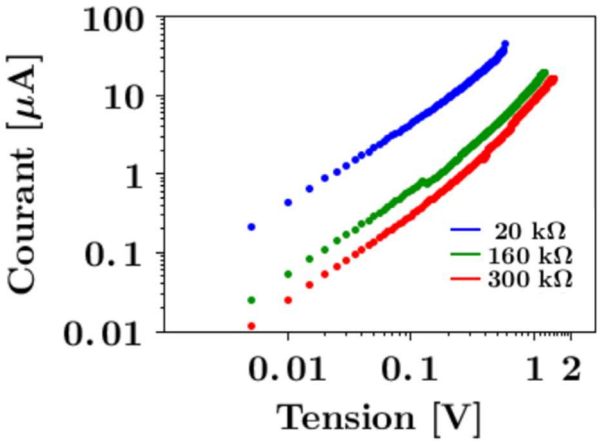


Figure III.7 : représentation des mesures en échelle log-log. Jusqu'à environ 150mV les trois courbes sont parallèles et linéaires. Après 150mV, seule la courbe bleue, de faible résistance, le demeure.

Ainsi la modélisation du courant HRS se fait en deux temps :

- A faible tension, le courant sera modélisé linéairement.
- A forte tension, le courant sera exprimé conformément à l'exponentielle de la racine carré de la tension.

En Figure III.8 nous avons représenté le fit obtenu pour 4 résistances HRS différentes (le fit est représenté en pointillé, alors que les valeurs expérimentales sont représentées en traits pleins). La même formule a été utilisée pour les 4 résistances. Ainsi, pour modéliser un courant, il nous suffit d'entrer la résistance de la cellule (c'est-à-dire la valeur de courant lue à 0,1V, à température ambiante) et le modèle créé la courbe I-V correspondante. Afin de modéliser au mieux les courbes, nous avons fait varier la valeur de β, en fonction de la résistance, via l'équation :

$$\beta = 4.03V^{-1/2} - \frac{R_0}{R}(4.03V^{-1/2} - 2.5V^{-1/2})$$

Même si le fit semble relativement correct, pour les quatre valeurs de résistances, on remarque néanmoins quelques irrégularités dans le courant mesuré expérimentalement, qui explique les légères divergences avec le fit (cf. Fig. III.9 (a)). Ces variations de courants peuvent correspondre à des mouvements de lacunes d'oxygènes sous l'effet du champ électrique appliqué, ou à du bruit RTN. En effet, en Figure III.9 (b) nous avons repris une figure de [22], où le phénomène de bruit RTN conduit à des modifications de courant similaires à celles que nous avons obtenues.

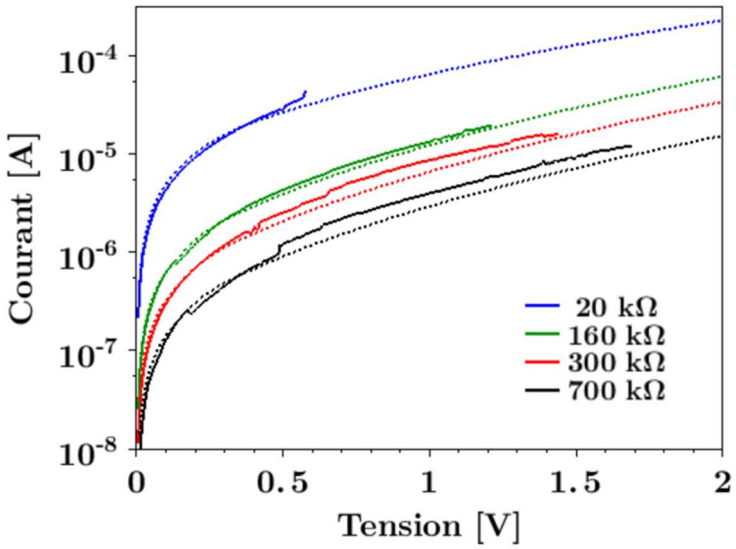


Figure III.8 : Modélisation du courant pour des résistances variées. Le fit est représenté en pointillé. Les traits pleins correspondent aux mesures expérimentales.

Nous reviendrons sur le problème du bruit RTN dans le dernier chapitre de ce manuscrit de thèse.

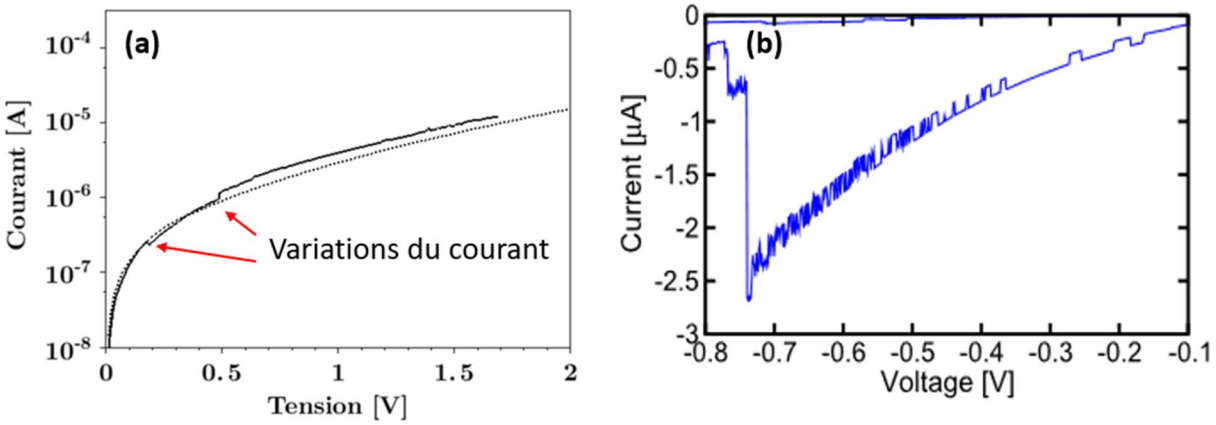


Figure III.9 : (a) Existence de variations de courants qui expliquent les divergences entre les courbes mesurées et les fits. (b) Courbe issue de [22] avec des oscillations en courants due à du bruit RTN.

2.4. Etude de la conduction de l'état LRS

La conduction dans l'état faiblement résistif est très simple, puisqu'elle est complètement ohmique : le courant est tout simplement proportionnel à la tension appliquée, conformément à la loi d'Ohm U=RI. Ceci a déjà été dit dans le premier chapitre de ce manuscrit et est maintenant considéré comme admis dans la littérature. En Figure III.10 nous avons brièvement tracé le retour de trois opérations de set à trois compliance différentes. On voit bien que le régime est linéaire. Le léger affaissement de la courbe bleue est dû au transistor en série avec l'OxRAM.

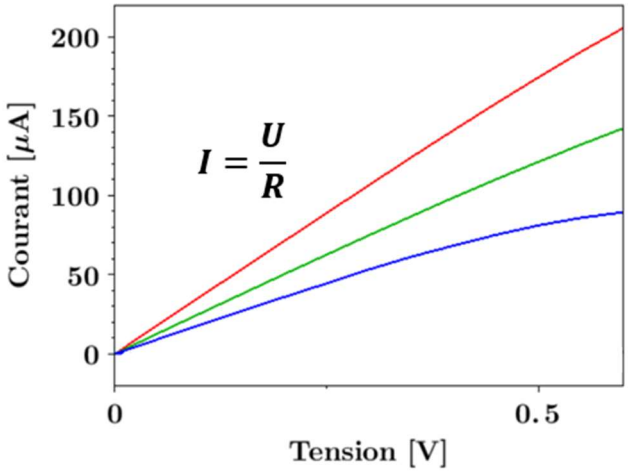


Figure III.10 : La conduction dans le régime LRS est ohmique. Les trois courbes correspondent à trois valeurs de compliance différentes.

2.5. Lien entre la résistance électrique et le nombre de lacunes d'oxygène

Pour l'instant, nous sommes capables de calculer le courant lié à une résistance fixe. Nous cherchons maintenant à relier cette résistance à un paramètre microscopique, qui sera le nombre de lacunes d'oxygène. Comme pour le calcul du courant, nous considérerons séparément les états HRS et LRS. La Figure III.5 nous indique la zone frontière entre les deux régimes (aux alentours de 10^4 Ω). Or cette valeur est extrêmement proche de la valeur de l'inverse du quantum de conductance G_0, de 12900Ω. Pour rappel, on observe pour certaines configurations (Quantum Point Contact) où l'échelle des chemins de conduction devient atomique, une quantification de la conductance. Celle-ci prend alors des valeurs discrètes, multiples de G_0. Or, ce phénomène a déjà été observé dans des dispositifs RRAM [23, 24, 25, 26] (cf. Fig. III.11).

Pour notre modèle nous supposerons donc que cette résistance de 12900Ω, qui sépare le régime métallique du régime Poole-Frenkel, correspond à l'état où la conduction est assurée par un filament monoatomique (cf. Fig. III.12). Pour les résistances plus élevée, on considérera que ce filament est partiellement rompu. Pour les résistances plus faibles, on considérera que ce filament est plus épais.

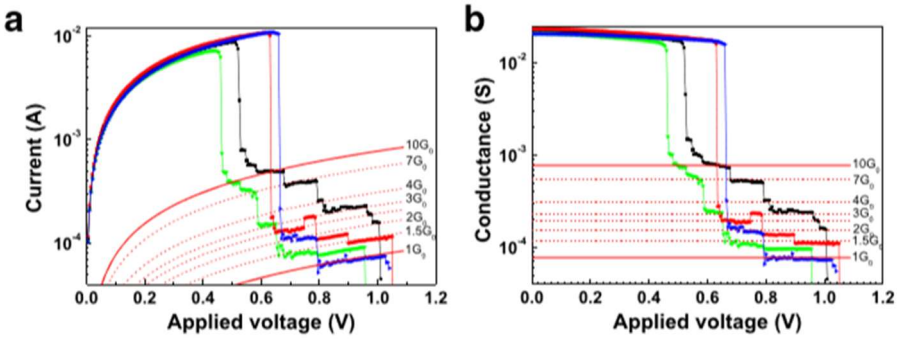


Figure III.11 : (a) Observation du phénomène de quantification du courant lors d'opération de reset sur des dispositifs de HfO2. (b) Courbes de conductance associée [23]

Il nous reste maintenant à définir comment, pour les deux régions, passer du nombre de lacunes d'oxygène à la résistance électrique.

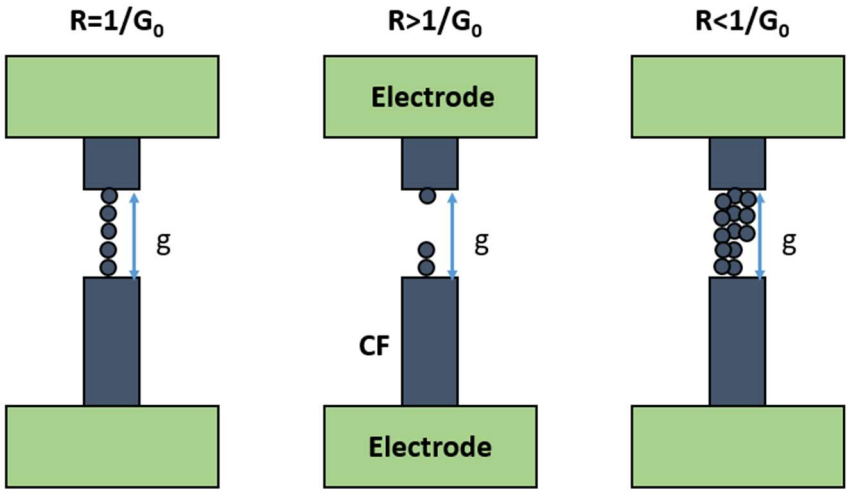


Figure III.12 : Schématisation de l'état du filament conducteur et de la zone de travail pour les trois états résistifs possibles.

2.5.1 Cas de l'état fortement résistif

L'état présentant une résistance supérieur à $1/G_0$ correspond à ce que nous appellerons ici l'état HRS. Ce régime a pour mode de conduction un régime assisté par pièges que nous avons décidé de modéliser par un modèle Poole-Frenkel. De façon très générale, la résistance est ici décrite par l'équation $\boldsymbol{R = A.e^{\frac{E_{AC}}{kT}}}$ (1), avec $\boldsymbol{E_{AC} = 0.030375.log_{10} R - 0.1256eV}$ (2).

Il nous reste à définir le terme A, en fonction des paramètres physiques qui nous intéressent, à savoir la distance g, et le nombre de lacune n, contenues dans ce gap g. Dans [27], S. Ambrogio et al. utilisent la formule suivante, également dans un contexte de modélisation via un courant Poole-Frenkel :

$\boldsymbol{R = B.\frac{g}{Nt}.e^{\frac{E_{AC}}{k.T}}}$ (3), avec B étant une constante et Nt le nombre de lacunes dans le gap g. Cette formule suppose que chaque lacune présente dans le gap participe au courant. Ainsi cette formule est assez simple puisque le courant Poole-Frenkel est directement proportionnel au nombre de lacunes, qui permettent le passage du courant.

Cette formule est donc parfaitement adaptée au contexte de notre modèle, c'est pourquoi nous l'utiliserons pour décrire la résistance dans l'état HRS. Il nous reste alors à calculer la constante B. Pour ce faire, nous allons procéder en recherchant la valeur permettant la continuité avec R=1/G0, correspondant au schéma de gauche de la Figure III.12. On cherche ainsi à résoudre l'équation :

$$\boldsymbol{R = \frac{1}{G_0} = B.\frac{g}{Nt}.e^{\frac{E_{AC}(R)}{k.T}}}$$

Dans cette configuration, le nombre de lacunes d'oxygène présentes est tout simplement égale à g/d, avec d étant la "taille" d'une lacune, c'est-à-dire, pour approximer, le double de la longueur d'une liaison Hf-O (qui est de 0.22nm [28]). Ainsi, on obtient une équation très simple pour B :

$$\boldsymbol{B = \frac{R_0}{d}e^{\frac{-E_{AC}(R_0)}{k.T}}} \quad (4)$$

Cette constante ne dépend donc que du matériau utilisé comme oxyde.

2.5.2 Cas de l'état faiblement résistif

Pour l'état LRS, nous avons choisi de rester sur le principe du quantum de conductance. Ainsi, l'idée de départ est de supposer une conductance multiple du quantum de conductance, c'est-à-dire une résistance de la forme : $\boldsymbol{R = \frac{1}{N_{path}.G_0}}$ (5), où N_{path} est le nombre de filaments monoatomiques qui relie les deux parties du filament conducteur.

N_{path} peut s'exprimer de la façon suivante :

$N_{path} = E(\frac{N_t}{\frac{g}{d}})$, E étant l'opérateur partie entière.

Cependant, expérimentalement, nous n'avons jamais observé de discrétisation de l'état LRS, n'autorisant que des valeurs de conductance multiples de G_0. Ceci peut s'expliquer par le fait que le transport est lié à différents chemins de conduction, plus ou moins complexes, qui viennent moduler l'effet de chaque atome, d'où la perte de la discrétisation. C'est pourquoi nous avons préféré passer cette expression en continue : $\boldsymbol{N_{path} = \frac{N_t}{\frac{g}{d}}}$ (6).

Pour ces formules, nous faisons l'approximation que les lacunes sont générées préférablement selon le même axe (cf. Fig. III.13). Physiquement on peut supposer que c'est la zone où le champ électrique est le plus fort, et donc où la probabilité de génération de lacunes est la plus forte.

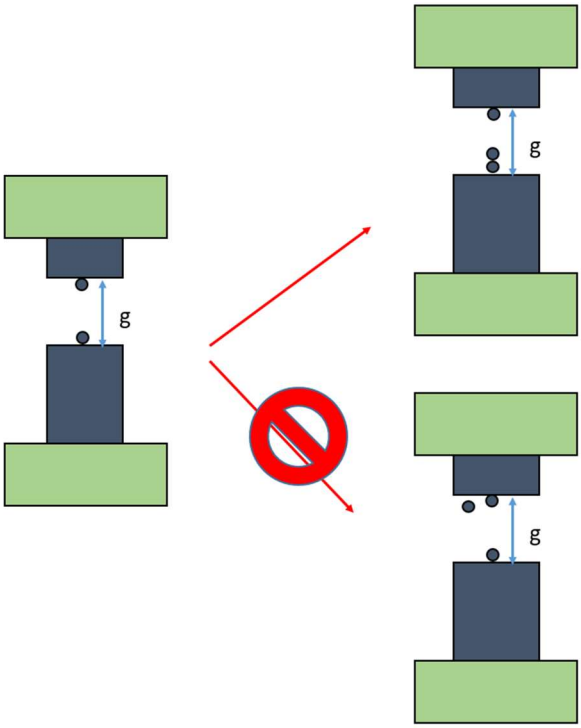


Figure III.13 : Les lacunes sont générées selon le même "chemin". Ainsi, dans l'exemple du schéma, les trois lacunes sont toutes sur la même ligne, ce qui justifie l'emploi de l'équation (6).

2.5.3 Tracé du nombre de lacunes en fonction de la résistance

On peut alors tracer, pour une valeur de g et une valeur de r (le rayon du filament conducteur), la courbe décrivant le nombre de lacunes présentes dans la zone de travail g, pour chaque valeur de résistance (cf. Fig. III.14).

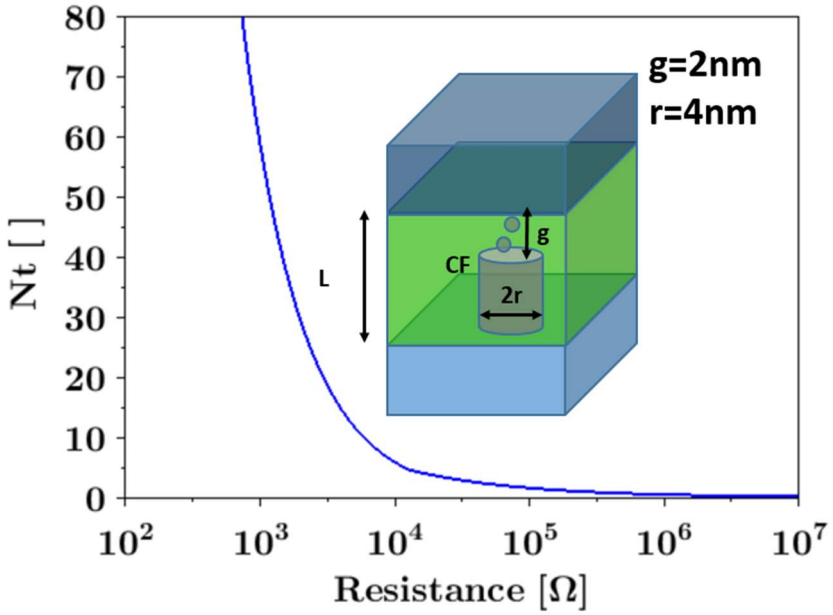


Figure III.14 : Lien entre le nombre de lacunes présentes dans le gap et la résistance électrique. Ici nous avons pris un gap de 2nm et un rayon de filament de 4nm.

Cette courbe est très importante pour la suite. Lorsque notre modèle fonctionne et génère/annihile des lacunes d'oxygène, il recalcule via cette courbe la résistance obtenue et peut ainsi en déduire le courant.

A noter que, dans cette version du modèle, Nt prend des valeurs continues. En fonction des besoins, il nous est tout à fait possible de choisir de ne prendre que des valeurs discrètes pour Nt.

2.6. Génération de lacunes d'oxygène

Comme nous l'avons déjà dit, l'opération de set se base sur la génération de lacunes d'oxygène au sein de l'oxyde métallique. Celles-ci vont augmenter la conduction dans le dispositif. Le mécanisme de la génération de lacunes d'oxygène a déjà été abordé dans une partie du premier chapitre de ce manuscrit de thèse.

Les travaux sur lesquels nous nous reposerons sont ceux de McPherson [28, 29, 30]. Dans ces papiers, l'auteur décrit le breakdown de diélectriques via l'action du champ électrique. Selon ce modèle, un champ électrique suffisamment intense a pour effet d'affaiblir les liaisons dans les diélectriques high-k, qui peuvent ainsi se rompre plus facilement, par processus de Boltzmann classique. Autrement dit, le champ électrique a pour effet de diminuer l'énergie d'activation nécessaire à la rupture d'une liaison Hf-O (cf. Fig. III.15).

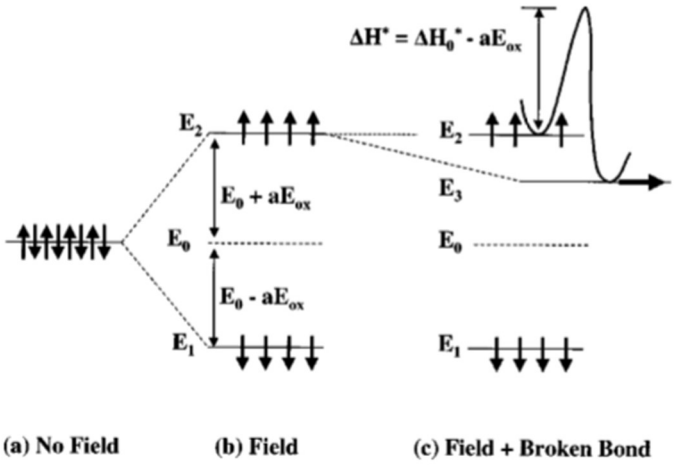


Figure III.15 : Effet du champ électrique sur la structure du diélectrique dans le cas de SiO_2 (a) Avant l'application du champ électrique, tous les dipôles ont la même énergie. (b) L'application d'un champ électrique entraine une différenciation des états énergétiques : les dipôles parallèles au champ électrique ont alors une énergie inférieure à ceux dans une configuration antiparallèle. (c) Afin d'aller vers un état énergétique plus faible (E_3), au prix d'une énergie d'activation ΔH*, certaines liaisons Si-O vont se rompre pour accéder à un état de dipôle orthogonal au champ électrique. [30]

L'énergie d'activation pour briser la liaison est donnée par la formule : $\Delta \boldsymbol{H}^* = \Delta \boldsymbol{H_0}^* - \boldsymbol{p_0} \frac{\boldsymbol{2+k}}{\boldsymbol{3}} \boldsymbol{E}$ [28], où ΔH_0* est l'énergie d'activation requise en l'absence de champ électrique, k est la constante diélectrique, E est le champ électrique et p_0 est la composante du moment dipolaire moléculaire opposé au champ électrique (sa valeur est de 4,4eÅ dans le cas de la liaison Hf-O [28].

En se basant sur ces travaux, on peut, à l'instar par exemple de l'équipe de l'Université de Modena [31], introduire un taux de génération de lacunes d'oxygène de la façon suivante :

$$\boldsymbol{G = \nu . \exp(-\frac{E_{AG} - b.F}{k_B.T})} \quad (7)$$

Avec υ étant la fréquence de vibration de la liaison Hf-O, E_{AG} l'énergie d'activation de génération (comparable à ΔH_0*), b=$p_0 \frac{2+k}{3}$, F étant le champ électrique (nous avons changé la notation afin d'éviter les confusions entre les symboles d'énergie et de champ électrique local), k_B la constante de Boltzmann et T la température locale.

Afin de rester le plus simple possible, nous avons choisi d'approximer le champ électrique comme étant simplement V/g. En effet, comme la zone de gap constitue la zone la plus résistive, il est raisonnable de penser que le champ électrique se concentre majoritairement dans cette zone, plutôt que dans la zone où le filament conducteur est intact.

On peut enfin déduire le nombre de lacunes d'oxygène, grâce à ce taux de génération de lacune, en intégrant l'équation (8) :

$$\boldsymbol{dNt = G(N_0 - Nt).dt} \quad (8)$$

Avec N_0 le nombre total de sites possibles pour les lacunes dans la zone de gap. N_0-Nt est donc le nombre de sites pouvant encore générer une lacune. N_0 dépend donc de g (taille du gap) et de r (le rayon du filament conducteur).

2.7. Recombinaison de lacunes d'oxygène

De la même façon que le set est lié à la génération de lacunes d'oxygène, le reset est lié à leur recombinaison avec les ions/atomes d'oxygène, et à la migration d'ions/atomes oxygène.

Notre modèle ayant pour but d'être analytique, nous avons besoin de modéliser ce phénomène via une équation simple. C'est pourquoi, comme pour le modèle de l'Université de Modena, nous utiliserons un taux de migration/recombinaison, de la même manière que nous avons utilisé un taux de génération pour le set.

Dans la littérature, on trouve des énergies d'activation de recombinaison faibles, de l'ordre de 0,2eV [31] (alors qu'à titre de comparaison les énergies d'activation de génération

sont généralement comprises entre 1 et 3eV [31, 32, 33]). Nous supposerons donc que l'évènement limitant n'est pas la recombinaison de l'ion oxygène avec la lacune, mais la migration de l'oxygène vers la zone où sont localisées les lacunes (le filament donc). Ainsi, nous nous contenterons d'introduire un taux de migration des oxygènes (d'une zone proche de l'électrode supérieure, vers la zone de travail (riche en lacunes d'oxygène, avant le reset)). La formule que nous utiliserons pour le reset sera donc la suivante :

$$\boldsymbol{M = \nu . \exp(-\frac{E_{AM} - k_D . F}{k_B . T})} \quad (9)$$

Avec E_{AM} l'énergie d'activation migration d'oxygène dans le matériau, k_D un facteur dépendant du milieu qui représente l'abaissement de cette énergie d'activation par le champ électrique.

L'évolution du nombre de lacunes est alors décrite par l'équation :

$$\boldsymbol{dNt = -M . Nt . dt} \quad (10)$$

Lors du reset (contrairement au set qui a besoin d'une compliance pour ne pas entraîner un breakdown complet du diélectrique) si l'on arrête d'augmenter la tension, l'opération s'arrête et on ne "reset" pas toute la cellule. En effet, comme nous l'avons vu dans le premier chapitre, à mesure que l'on détruit le filament, on abaisse la tension qui s'applique dans la zone riche en lacune (puisque la majeure partie de la tension retombe sur les zones résistives). Il faut donc augmenter la tension totale pour poursuivre l'opération. Nous avons donc introduit dans le modèle des facteurs de conductivité pour différencier la conductivité de l'oxyde "intact" par rapport à l'oxyde en présence de lacunes. Ces facteurs ne sont à priori pas connus et seront des paramètres de fit, lorsque nous voudrons modéliser des courbes expérimentales.

2.8. Calcul de la température

Le calcul de la température est une partie importante du modèle. En effet, tous les phénomènes mis en jeu (génération de lacunes, migration d'ions oxygène notamment) sont activés en température.

Nous avons au début choisi de modéliser l'évolution de la température via un effet Joule classique. Cependant, nous n'arrivions ainsi pas à reproduire correctement les courbes d'évolution de la tension de set en fonction de la vitesse de rampe (cf. Chapitre II), pour différentes valeurs de températures. En effet, pour une température de 250°C, notre modèle prévoyait alors un switching à 0.75V (à une vitesse de rampe de 10^6 V/s), alors que la valeur expérimentale est supérieure à 0.9V.

C'est pourquoi nous avons opté pour le concept de température électronique [34, 35]. C'est de plus, plus cohérent d'un point de vue physique : le phénomène de switching est extrêmement local et on ne peut pas vraiment parler d'une température dans le vrai sens du terme, mais plus d'un chauffage de porteurs dans une région très confinée. En effet nous sommes ici dans le cas où un gap sépare deux conducteurs mésoscopiques (les deux portions de filament). Il a été montré dans [35] que dans ce cas, la température électronique locale suit l'équation (11). Dans [36], S. Blonkowski et T. Cabout utilisent également le concept de température électronique pour modéliser la température au sein de la zone de rupture du filament.

$$\boldsymbol{T_e(T, V_{ox}) = \frac{T}{1+\alpha} + \frac{\frac{q.V_{ox}}{2.k_B}.\coth\left(\frac{q.V_{ox}}{2.k_BT}\right)}{1+\alpha^{-1}}}$$, (11) avec Vox la tension aux bornes de l'oxyde

Avec T étant la température extérieure, V_{ox} la tension aux borne de la partie encore oxydé (donc ici la zone de travail g) et α le facteur de couplage électron-phonon [36].

3. Fonctionnement du modèle

Nous venons de présenter les différents mécanismes mis en jeu dans notre modèle. Nous allons maintenant expliquer son fonctionnement global et comment il s'articule.

3.1. Schéma de fonctionnement

La Figure III.16 décrit le fonctionnement de notre modèle.

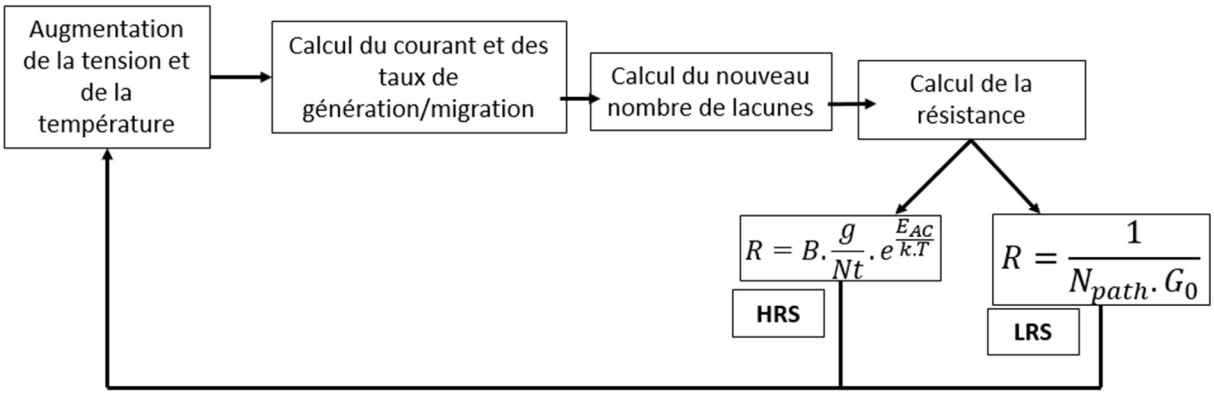


Figure III.16 : Schéma de fonctionnement du modèle.

3.2. Impact de paramètres d'entrée

Parmi les paramètres du modèle, un grand nombre ne peut pas être fixé avant de tenter de fitter des courbes expérimentales. Dans le cas de l'opération de set, les principaux paramètres sur lesquels nous n'avons qu'une faible idée de l'ordre de grandeur sont la taille de la zone de gap g, l'énergie d'activation de génération de lacune et le rayon du filament.

Le modèle a pour valeurs de sortie à la fois la résistance obtenue après opération, ainsi que la tension à laquelle le set/reset se produit. La résistance initiale est une valeur d'entrée. Avant la simulation, le programme calcule le nombre de lacunes présentes pour obtenir la résistance souhaitée, avec la valeur de gap rentrée.

Dans cette partie nous montrerons l'impact des principaux paramètres de fit sur ces valeurs de sortie. Par la suite, nous nous servirons des résultats obtenus dans le chapitre précédent pour fixer certains de ces paramètres.

En Figure III.17, nous avons représenté l'influence d'un changement de valeur de l'énergie d'activation de génération de lacunes pendant l'opération de set, sur les évolutions du courant et de la résistance de la cellule tout au long de l'opération. Ces courbes ont été obtenues avec un rayon de 1nm, une compliance de 200µA, un gap de 2nm, une résistance initiale de 230kΩ, une tension maximale de 1.5V et un temps de montée de 1µs. Le passage de 1.5eV à 3eV de l'énergie d'activation change radicalement les courbes.

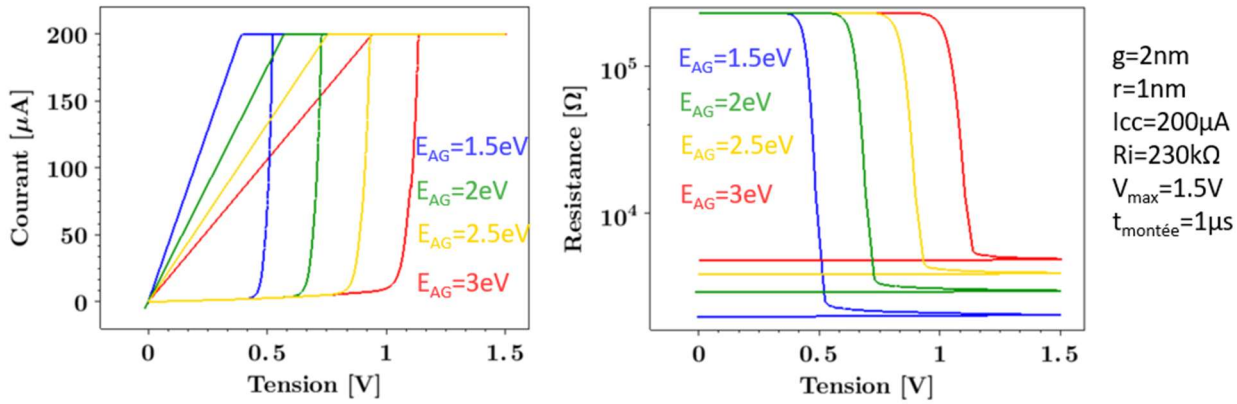


Figure III.17 : Influence de l'énergie d'activation de génération de lacune durant l'étape de set.

En toute logique, augmenter l'énergie d'activation augmente la tension de set (il faut un champ électrique plus important pour générer les lacunes). On génère ainsi moins de lacunes, ce qui nous amène à obtenir une résistance LRS finale plus élevée.

La Figure III.18 représente l'impact du changement de la valeur de g. Nous avons pris ici une valeur d'énergie d'activation de génération de lacunes de 2eV. Les autres paramètres

n'ont pas été modifiés, par rapport aux courbes précédentes. Les valeurs prises pour la taille de la zone de gap sont 1nm et 3nm.

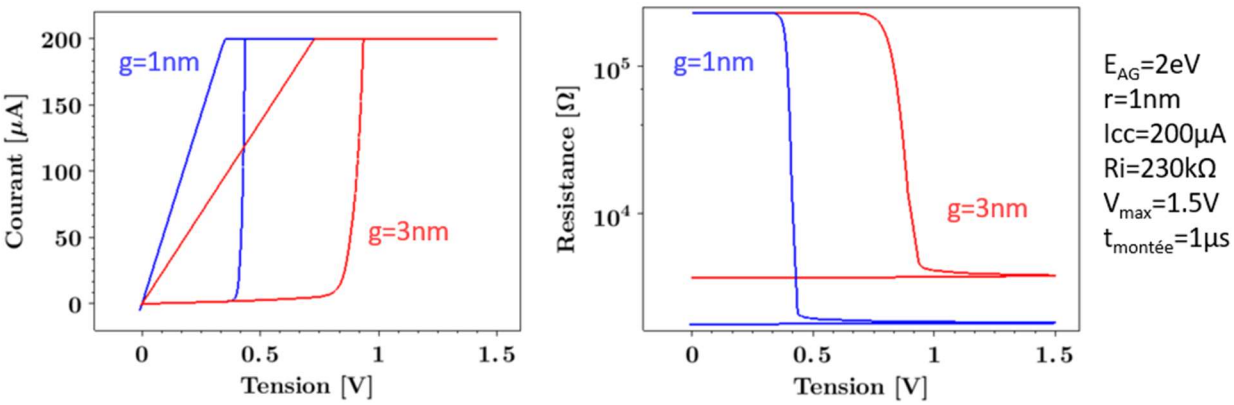


Figure III.18 : Influence de la taille de la zone de gap durant l'étape de set.

Comme l'énergie d'activation de génération, la taille du gap a une importance cruciale (puisque le champ électrique, contenu en exponentiel du terme de génération de lacune, dépend de g). Un gap plus élevé revient à un champ électrique plus faible dans la zone de gap, ce qui limite la génération de lacunes, d'où une tension de set plus élevée et une résistance de l'état LRS final plus élevée.

Enfin, nous avons représenté en Figure III.19 l'influence du rayon du filament. Pour ces courbes nous avons pris un gap de 2nm et une énergie d'activation de génération de 2eV. Les autres paramètres sont restés inchangés. Les deux valeurs de rayon de filament conducteur sont 1nm et 6nm.

Même si elle est loin d'être négligeable, l'influence du rayon du filament (varié ici d'un facteur 6) est secondaire, relativement à celle de l'énergie d'activation de génération et du gap. En effet, r ne figure pas dans l'expression du taux de génération. Si un filament large permet un switching à une tension plus faible c'est simplement car il augmente le nombre de "places disponibles" pour les lacunes d'oxygène, et augmente donc de façon linéaire la probabilité de génération de lacunes.

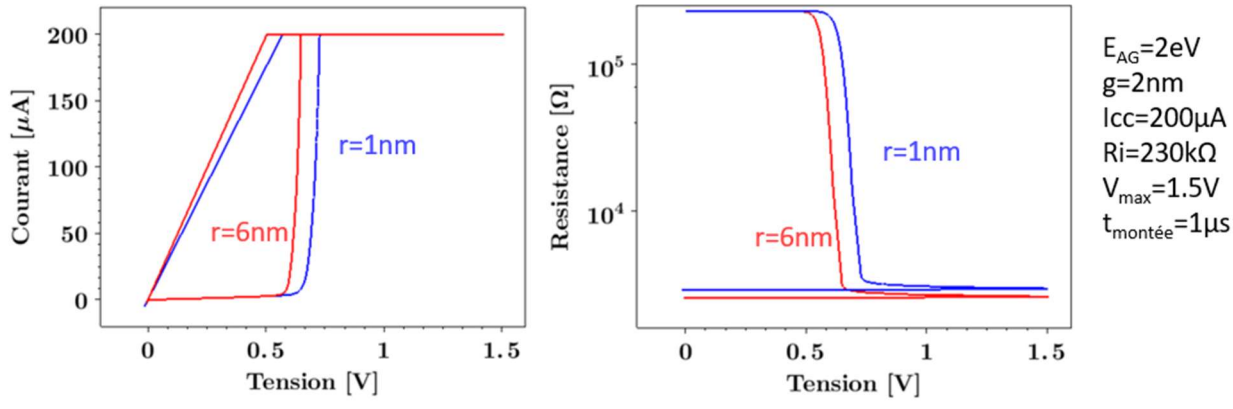


Figure III.19 : Influence du rayon du filament conducteur durant l'étape de set.

Le modèle s'appuie donc sur un certains nombres de paramètres physiques. Si le gap et le rayon du filament sont censés pouvoir évoluer en fonction des paramètres électriques (tension, courant de compliance) de set et reset, l'énergie d'activation de génération est une constante physique qui doit rester la même pour une association de matériaux donnée (par exemple, des OxRAM de HfO_2 avec une électrode supérieure en titane). La partie qui suit va donc consister à déterminer cette énergie d'activation.

4. Etalonnage du modèle

Cette étape de détermination des paramètres fixes du modèle peut être qualifiée d'étalonnage du modèle. Nous commencerons par expliquer la démarche via laquelle nous allons déterminer la valeur de l'énergie d'activation de génération pour des dispositifs d'oxyde de hafnium avec électrode supérieure en titane (puisque nous avons montré que l'électrode supérieure jouait un rôle sur les valeurs de tension de switching).

4.1. Détermination de l'énergie d'activation de génération pour HfO_2/Ti

Dans le chapitre précédent, nous avons caractérisé l'influence de la vitesse de rampe sur les valeurs de tension de set et de reset (cf. Fig. II.26).

Nous allons donc tenter de reproduire cette courbe avec notre modèle. Dans la mesure où les paramètres utilisés pour ces mesures étaient fixes (hormis la vitesse de rampe), nous savons que le gap et le rayon n'ont pas de raisons d'être modifiés. Nous devons donc reproduire cette courbe en trouvant un jeu de paramètre (E_{AG}, g et r) constant.

De plus nous disposons des courbes pour les mêmes dispositifs, mais à différentes températures. Ainsi, ce jeu de paramètres doit également être capable de fitter ces courbes (cf. Fig. II.27).

Comme r n'a qu'une influence faible sur les valeurs de tension de set, nous avons décidé de le laisser constant à 1nm pour n'avoir plus que 2 variables.

De cette façon nous n'avons trouvé qu'un seul jeu de paramètres permettant de fitter les courbes. Ce jeu de paramètres correspond à une énergie d'activation de 2.7eV et un gap de 2.3nm.

En Figure III.20, nous avons représenté les courbes de courant-tension pour 6 vitesses de montée différentes (10^3, 10^4, 10^5, 10^6, 10^7 et $10^8 V.s^{-1}$).

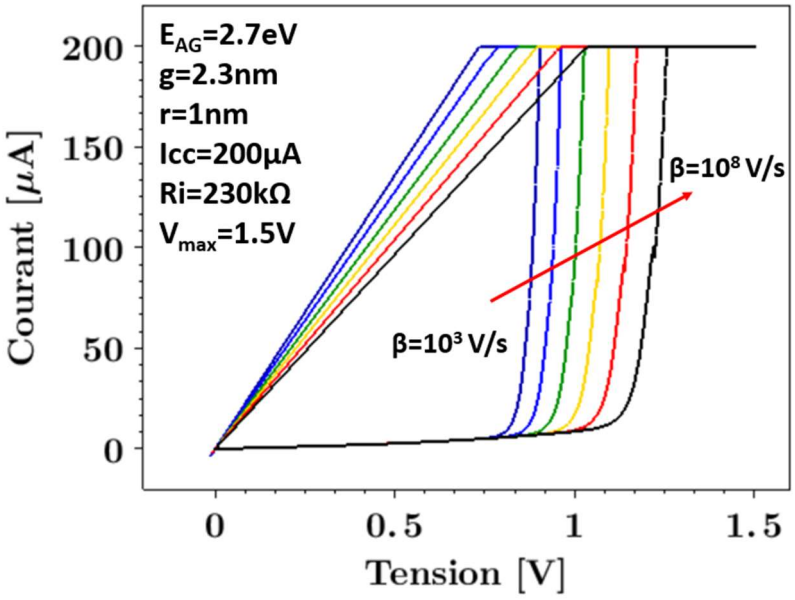


Figure III.20 : Courbes courant-tension pour différentes valeurs de β.

De même que sur les mesures expérimentales, on observe bien une augmentation de la tension de set, à mesure que l'on augmente la vitesse de rampe. De plus, on remarque que la tension de set augmente de façon régulière à mesure que l'on augmente β. La courbe d'évolution de la tension de set en fonction de β est bien une droite.

En Figure III.21 nous avons donc comparé les mesures expérimentales, aux résultats obtenus avec le jeu de paramètre adapté (en pointillé).

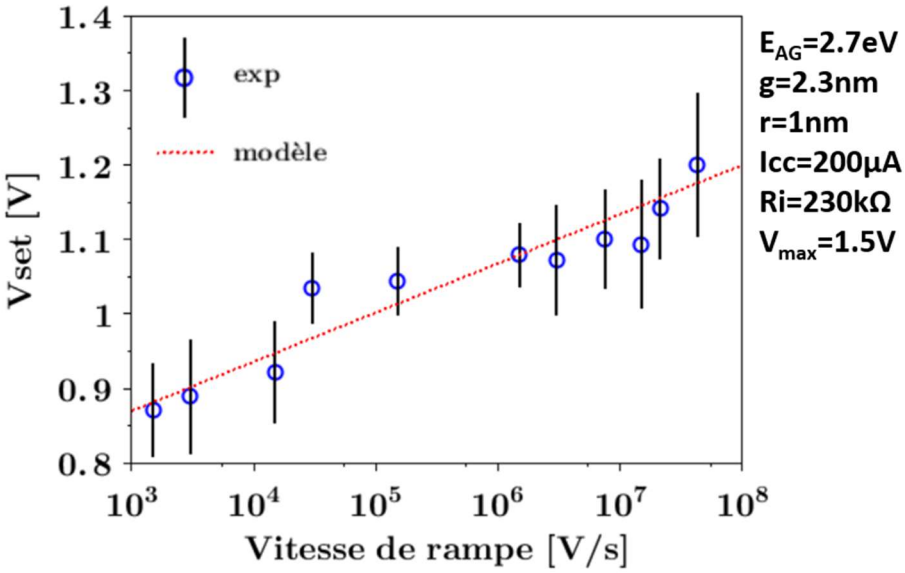


Figure III. 21 : Comparaison entre les mesures expérimentale et les résultats du modèle de l'évolution de la tension de set en fonction de la vitesse de rampe.

Pour confirmer la validité de notre jeu de paramètre, nous devons également vérifier si notre modèle reproduit bien les mêmes mesures, pour différentes températures. Nous avons donc repris les mesures réalisées en température dans le deuxième chapitre de ce manuscrit de thèse et avons regardé si le modèle reproduisait correctement l'activation en température. Les courbes associées à ces mesures sont représentées en Figure III.22.

Sur cette courbe, dans un souci de lisibilité nous n'avons représenté que 3 températures (25, 200 et 250°C), sur les 5 testées. Sans avoir à changer un seul paramètre (hormis, logiquement, la température ambiante), on constate que le modèle prévoit de façon très efficace l'activation thermique de l'opération de set, ce qui valide la formule de la température électronique développée plus tôt dans cette partie.

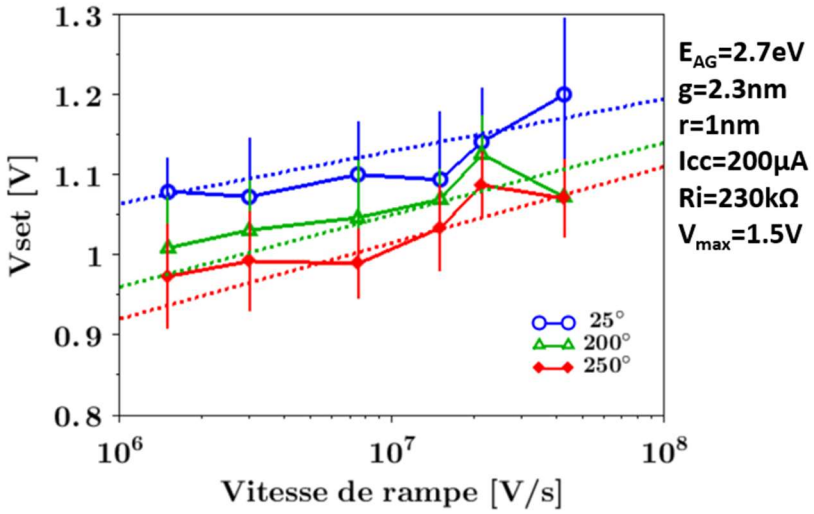


Figure III.22 : Comparaison modèle/expérience de l'activation en température du phénomène de set.

Grâce à ces mesures, on peut maintenant considérer notre modèle comme opérationnel, sur les dispositifs de HfO_2/Ti.

4.2. Fit en quasi-statique du set

Pour l'instant, comme en attestent les courbes de la Figure III.19, nous avons implémenté la compliance de la façon la plus simple possible : lorsque le courant atteint la valeur de compliance fixée, la tension aux bornes de la mémoire chute et on obtient un diviseur de tension entre la mémoire et le transistor.

Dans la réalité, le transistor n'est pas parfait et toute la tension ne retombe pas sur l'OxRAM avant que la compliance ne soit atteinte. Ainsi, afin de mieux modéliser les courbes courant-tension en quasi-statique, nous avons rajouté dans le modèle les caractéristiques des transistors utilisés.

On obtient alors des fits de courbes quasi-statiques assez satisfaisants (cf. Fig. III.23). En effet, pour les 3 compliances utilisées, le moment du set, ainsi que la résistance après switching sont très correctement reproduits par le modèle (en traits pleins). A noter qu'ici encore, nous avons utilisé une valeur de gap de 2.3 nm, une énergie d'activation de génération de lacunes d'oxygène de 2.7eV et un rayon de filament de 1nm.

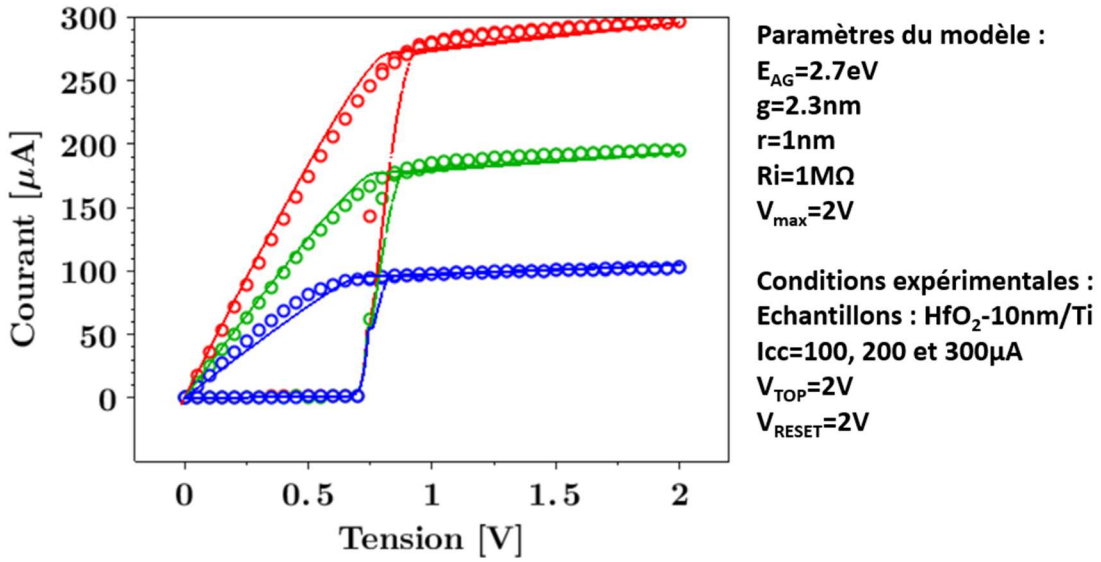


Figure III.23 : Fit de 3 opérations de set en quasi-statique, pour trois tensions de grilles différentes.

4.3. Opération de reset

De même que pour l'opération de set, nous avons essayé de reproduire l'impact de la vitesse de rampe sur la tension de reset, avec le modèle. De même que pour le set, nous pouvons observer, sur la Figure III.24 que la tension de reset augmente de manière régulière lorsque l'on augmente β. La courbe d'évolution de la tension de reset en fonction de β est encore bien une droite.

Ces courbes ont été obtenues avec une valeur de gap identique à celle utilisée pendant le set, soit 2.3nm, un rayon de filament de 1nm, une résistance initiale de 1000Ω, une énergie d'activation de migration/recombinaison de 1.65eV et une valeur de k_D de 55eÅ. En revanche, contrairement au set (où p_0 est connu), kD n'est pas connu. Nous avons donc ici plusieurs jeux de paramètres possibles aboutissant aux mêmes courbes. Nous ne pouvons donc pas comparer la valeur d'énergie d'activation que nous avons utilisée avec d'autres études.

Augmenter la vitesse de rampe augmente la tension à laquelle la cellule reset, et donc le courant maximal de l'opération. Ici, le courant est relativement élevé car nous avons voulu comparer nos résultats avec ceux obtenus dans le chapitre II, lorsque la résistance de l'état LRS était très faible.

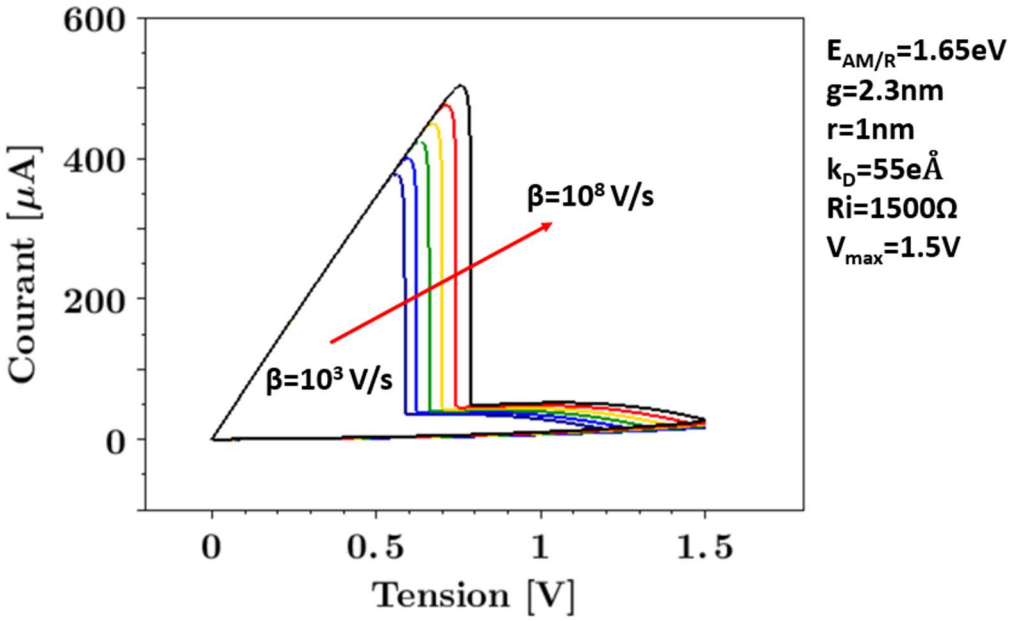


Figure III.24 : Courbes courant-tension pour différentes valeurs de β.

Pour ces courbes, nous avons pris en compte l'influence du transistor, puisque, comme nous l'avons étudié dans le chapitre II, une partie non négligeable de la tension totale appliquée retombe sur le transistor, malgré l'emploi d'une très forte tension de grille.

Nous avons donc ensuite pu comparer la courbe expérimentale obtenue dans le chapitre II avec ces résultats. Cette comparaison est représentée en Figure III.25.

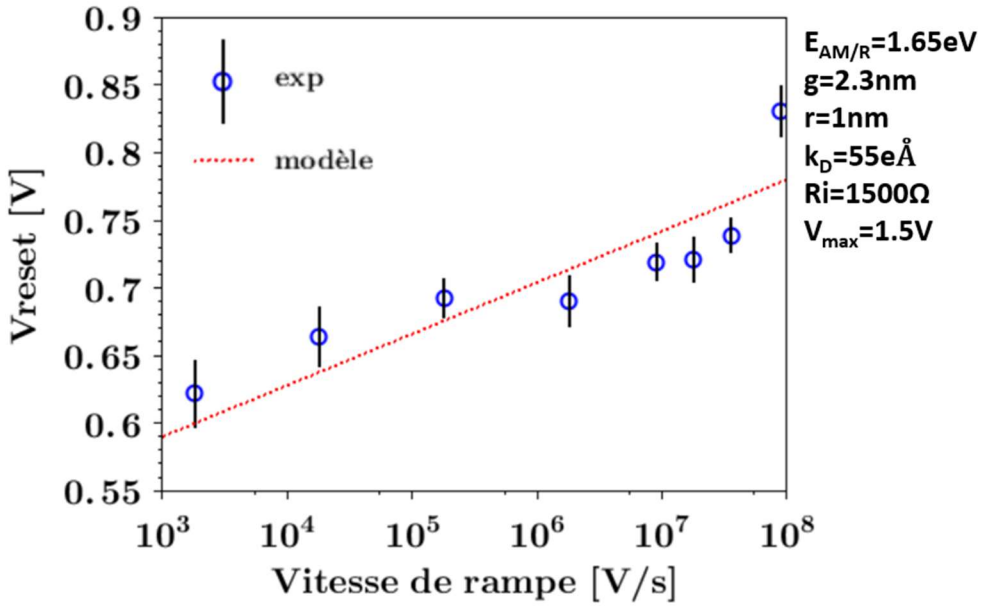


Figure III.25 : Comparaison entre les mesures expérimentale et les résultats du modèle de l'évolution de la tension de set en fonction de la vitesse de rampe.

En conclusion, que ce soit pour le set, ou le reset, le modèle capte de façon efficace l'influence de la vitesse de la rampe sur les tensions de switching, et le fit de ces courbes par le modèle permet de déterminer les paramètres du modèle.

4.4. Mode discret

Les courbes courant-tension de reset, présentée en Figure III.24 sont très "lisses". En effet, à mesure que le reset avance, le nombre de lacunes diminue, mais de manière continue, comme représenté en Figure III.26. Ainsi, après la première brusque chute de courant (associée au moment du reset), il n'y a donc pas de saut en courant, comme on a l'habitude d'en observer sur les courbes quasi-statiques.

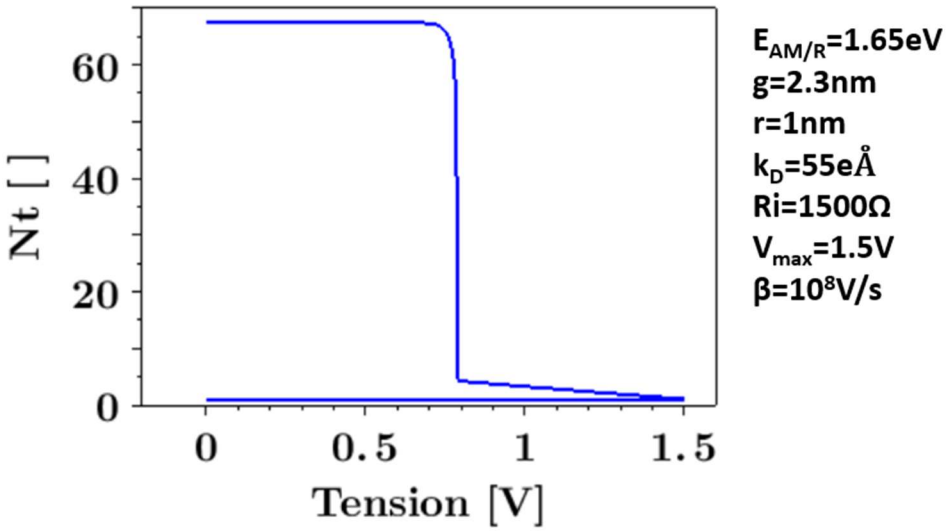


Figure III.26 : Evolution du nombre de lacunes d'oxygène pendant le reset. Cette courbe correspond à celle obtenue avec une vitesse de rampe de 10^8, de la Figure III.24.

C'est la raison pour laquelle nous avons également implémenté un mode "discret" dans le modèle. Ce mode consiste simplement à ne prendre en compte que la partie entière du nombre Nt. Cela revient à sommer les chemins de conductions. Avec ce mode, les courbes courant-tension obtenue sont plus en accord avec ce que l'on obtient expérimentalement sur les mesures en quasi-statique. Dès qu'une lacune disparaît on observe bien une chute de courant, correspondant à la baisse de conductivité engendrée par la perte de cette lacune d'oxygène. Notons cependant que cela ne change pas les résultats obtenus ci-dessus : la courbe prise en exemple en Figure III.27 (a) correspond à la courbe noire de la Figure III.24 et on voit bien que le moment du reset est inchangé. On remarque facilement les chutes de courant liés aux

moments où une lacune d'oxygène se recombine avec un ion oxygène. En figure III.27 (b) nous avons repris une courbe de [27], présentant les mêmes caractéristiques.

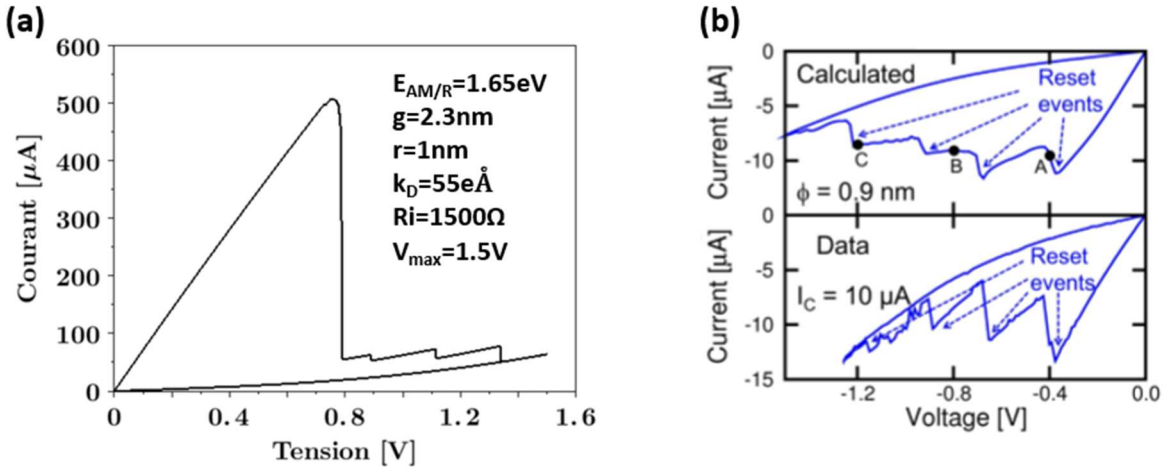


Figure III.27 : (a) courbe courant-tension d'un reset obtenue en mode "discret". (b) Allure similaire de reset, présenté dans [27].

4.5. Fit en quasi-statique du reset

L'utilisation du mode discret permet de fitter plus efficacement les courbes courant-tension en quasi-statique de l'opération de reset. En Figure III.28, nous avons fitté une courbe de reset, en quasi-statique (temps de montée de 2ms).

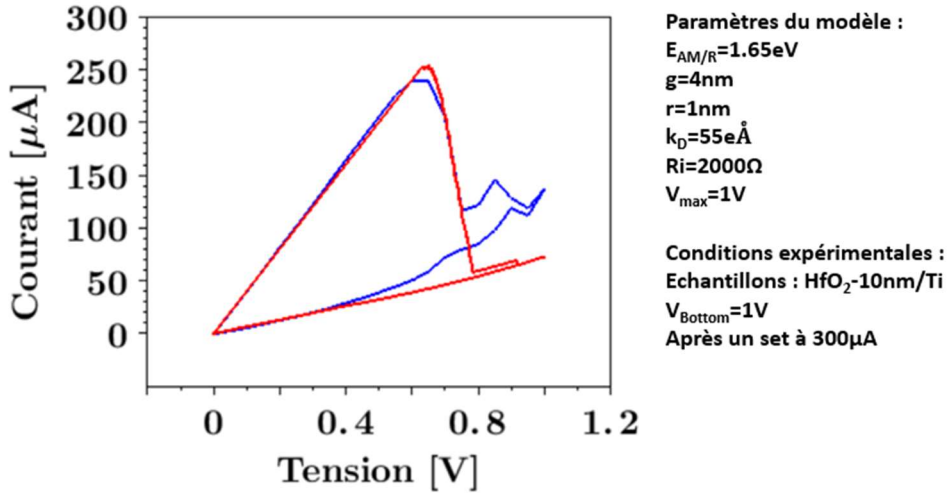


Figure III.28 : Fit d'une opération de reset en quasi-statique. Le fit est représenté en rouge tandis que la mesure est en bleu.

Si l'on ne reproduit pas parfaitement la courbe, sur les zones à fortes tensions notamment, on obtient néanmoins exactement la même tension de reset ainsi que la même résistance HRS finale (puisqu'à faible tension, après le reset, le fit et la mesure sont superposés). A noter également qu'il est très compliqué pour un modèle de fitter correctement toutes les

instabilités en courant sur les zones à fortes tensions d'une courbe de reset. En effet à cette tension, il est fort possible que des phénomènes de génération, migration et recombinaison se produisent à plus forte fréquence, de façon stochastique. Il faut également prendre en compte la très forte chance d'y observer de plus du bruit RTN, absolument pas pris en compte dans ce modèle, lié à la charge et décharge de défaut à proximité du filament.

4.6. Tests électriques avec différentes électrodes supérieures

Comme nous l'avons vu dans le premier chapitre, dédié à l'étude bibliographique sur les mémoires OxRAM, le matériau utilisé en tant qu'électrode supérieure joue un rôle important [20, 37, 38]. Or nous disposons de dispositifs constitués de différents matériaux d'électrodes supérieures (Ti, Ta, Pt, Zr et un matériau particulier dont nous ne préciserons pas la composition complète $CuTe_2X$, pour raison de confidentialité).

Nous avons décidé de mener une campagne de mesures assez lourde, visant à reproduire la courbe de l'influence de la vitesse de rampe sur la tension de set, grâce à notre véhicule de test doué d'un plot de mesure entre le drain du transistor et le dispositif mémoire, pour les 5 matériaux disponibles (cf. Fig. III.29).

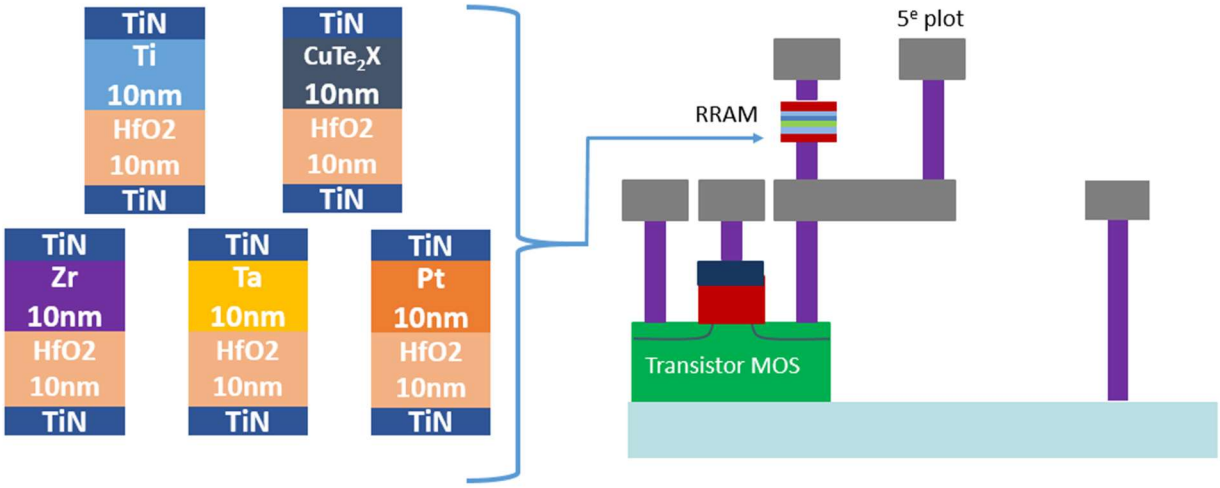


Figure III.29 : Dispositifs testés dans cette partie.

Le protocole expérimental est intégralement identique à celui suivi dans la partie 4 du second chapitre de ce manuscrit.

Les mesures avec l'électrode en titane ont déjà été réalisées dans le chapitre précédent. Pour le cas de l'électrode en platine, nous avons directement pris les résultats présentés dans [38] réalisés au sein du LETI.

De même que précédemment, nous avons ensuite utilisé notre modèle pour fitter les courbes d'évolution de la tension de set en fonction de la vitesse de rampe, pour les 5 stacks différents. Ces résultats sont présentés en Figure III.30. Nous avons également pu réaliser des simulations avec le logiciel commercial Ginestra™ [39], dans le cas de l'électrode de Ti. Ce logiciel se base sur le modèle physique mis au point par l'équipe de l'Université de Modena, dont nous avons déjà présenté un certain nombre de publications. Nous avons donc ajouté les résultats obtenus avec ce logiciel.

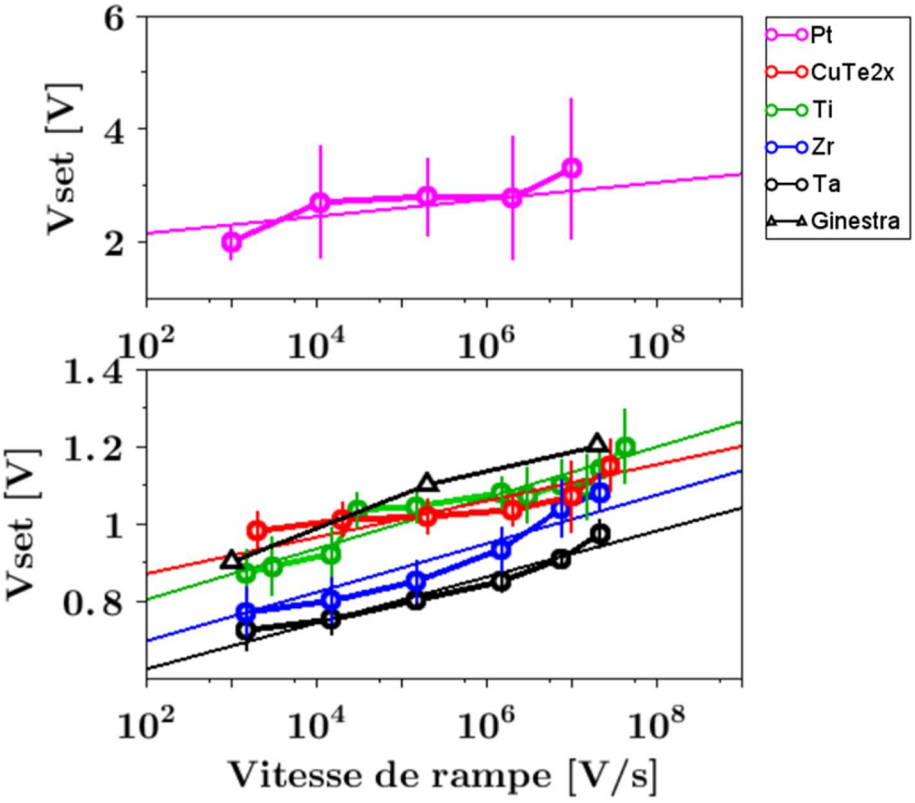


Figure III.30 : Evolution de la tension de set en fonction de la vitesse de rampe pour les 5 différents stacks et la simulation Ginestra™, ainsi que les fits obtenus avec notre modèle.

On peut tout d'abord constater la grosse différence entre les dispositifs avec l'électrode en platine et les autres. On voit ici encore l'intérêt d'utiliser une électrode active, plutôt qu'une électrode inerte, afin d'abaisser grandement les tensions mises en jeu, ainsi que l'influence de la vitesse de rampe : sur les temps vraiment courts, les tensions de set avec platine sont supérieures à 3V, alors qu'avec les autres, elles restent inférieures à 1.4V.

De plus, les courbes correspondant aux dispositifs avec l'électrode à base de Cu (en rouge) présentent une pente moins élevée, ce qui indique une plus faible influence de la vitesse de rampe sur la tension de switching. On voit ainsi qu'à très forte vitesse de rampe (temps de

montée très courts), la tension de set pour ces dispositifs devient plus intéressante qu'avec les dispositifs avec du Ti, alors qu'à faible vitesse, ces derniers semblent plus favorables.

Selon les courbes, nous avons fait varier l'énergie d'activation de génération de lacunes d'oxygène et parfois la valeur du gap. Les valeurs sont présentées dans la partie suivante.

4.7. Extraction des paramètres physiques et comparaison avec des simulations *ab initio*

Les valeurs d'énergie d'activation de génération de lacunes d'oxygène ainsi que du gap sont résumées pour chaque électrode supérieure dans le Tableau III.1. Comme nous l'avons dit précédemment, l'énergie d'activation représente l'énergie nécessaire pour générer une lacune d'oxygène. Nous avons ensuite contacté l'équipe de modélisation physique travaillant sur les OxRAM au sein du LETI pour savoir s'ils pouvaient calculer en *ab initio* ces énergies. Plus de détails sur la physique à l'origine de ces simulations sont disponibles dans [40].

TE	Modèle [eV]	*ab initio* [eV]	Gap [nm]	Type de défauts
Ti	2,7	2,63	2,3	Lacune d'O + O interstitiel dans Ti
Pt	4,4	4,9	3,7	Lacune d'O + O interstitiel dans HfO2
Zr	2,43	2,57	2,3	Lacune d'O + O interstitiel dans Zr
Ta	2,23	2,2	2,3	Lacune d'O + O interstitiel dans HfO2 [41]
$CuTe_2x$	3,8	3,4	1,3	Enthalpie de formation d'un ion Cu dans HfO_2 [42]

Tableau III.1 : Valeurs des énergies d'activation et de gap calculées avec notre modèle. Comparaison avec des calculs *ab initio*, et type de défauts générés dans chaque cas.

Dans le cas de Ti et de Zr, l'énergie calculée en ab initio correspond à l'enthalpie de formation d'une lacune d'oxygène au sein de HfO_2 et de la formation d'un oxygène interstitiel dans l'électrode supérieure (cf. Figure III.31). On constate que ces deux valeurs sont assez proches et correspondent bien aux valeurs calculées par notre modèle. A noter que cette énergie calculée en *ab initio* est ici purement thermodynamique et ne prend pas en compte une éventuelle barrière d'énergie (qui est très délicate à calculer dans le cas d'une interface).

Pour l'électrode en platine, il n'est pas favorable du tout de placer un oxygène interstitiel au sein du platine (9.58eV). Ici l'énergie obtenue de 4.9eV correspond à la formation d'une paire de Frenkel au sein de HfO_2.

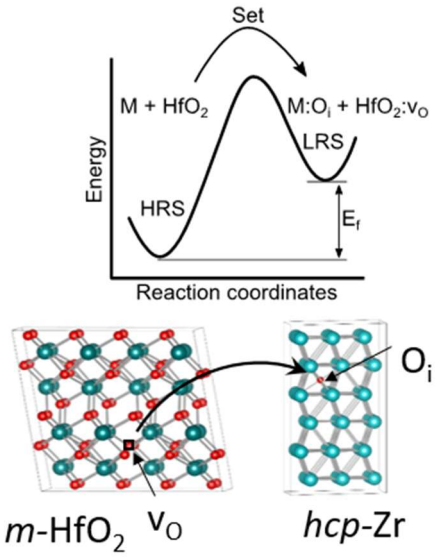


Figure III.31 : Schéma de la simulation *ab initio*, dans le cas d'une électrode en Zr, illustrant l'échange d'oxygène entre la matrice de HfO_2 et l'électrode supérieure. Les valeurs calculées en *ab initio* correspondent à l'énergie E_f.

Pour l'électrode en tantale, nous avons dû nous référer à une publication de l'IMEC [41] pour trouver une valeur proche de celle que nous avons obtenue (nous trouvions des enthalpies de formation trop faibles pour la formation d'un oxygène interstitiel dans Ta). Cependant cette valeur semble correspondre à la formation d'un oxygène interstitiel dans HfO_2, comme pour l'électrode en platine. On comprend donc difficilement comment cette valeur peut être si faible comparé aux autres. Comme cette valeur est issue d'un papier externe au CEA, nous ne connaissons pas en détail la façon dont les calculs ont été menés. On peut supposer que la valeur déduite par notre modèle correspond bien à la formation d'un oxygène interstitiel dans Ta, mais qu'il y a ici une barrière énergétique plus importante, qui explique que l'on obtienne une valeur relativement élevée de 2.23eV.

Enfin, pour l'électrode à base de cuivre, le calcul de l'enthalpie de formation pour créer une lacune d'oxygène donne des résultats compris entre 5,05 et 6,14eV ce qui signifie que ce n'est pas une réaction favorable du tout. En revanche, le calcul de l'enthalpie de formation pour placer un atome de cuivre au sein de la matrice de HfO2 donne une valeur de 3,4eV [42], assez proche de ce que notre modèle prévoit. Ceci révèle un mécanisme de set complètement différent, ne reposant pas sur la génération de lacunes d'oxygène, mais la migration d'éléments métalliques de l'électrode supérieure vers l'oxyde, à la façon des CbRAM [43, 44]. On a donc ici une technologie basé sur du HfO_2, emblématique des OxRAM, mais hybride CbRAM puisqu'il semble que la conduction ne repose, du moins pas à 100%, sur les lacunes d'oxygène.

C'est d'ailleurs la conclusion que tirent Nail et al. dans [42] : ils démontrent que le filament est constitué de Cu ayant migré dans le diélectrique.

5. Conclusion

A l'opposé du second chapitre, intégralement expérimental, ce troisième chapitre du manuscrit de thèse se centrait sur la mise au point d'un modèle physique et semi-analytique. La réalisation de ce modèle a constitué une part importante du travail réalisé durant ces trois ans de thèse. Ce modèle utilise comme variable d'état le nombre de lacunes d'oxygène (ou autres types de défaut, en fonction du mécanisme de switching, qui dépend, comme nous l'avons vu des matériaux utilisés) au sein d'une zone dite de gap, qui représente la zone du filament conducteur qui est amenée à être modifiée par les opérations de set et de reset. Ces opérations sont liées à la génération, migration et recombinaison de défauts au sein du matériau diélectrique.

Ce modèle nous permet de modéliser le comportement des OxRAM sur les opérations d'écriture et d'effacement, à la fois en quasi-statique et en dynamique. Nous avons ainsi pu reproduire les résultats du chapitre II de ce manuscrit sur l'impact de la vitesse de rampe sur la tension de set. Ce faisant nous avons pu déduire des paramètres physiques microscopiques comme l'énergie d'activation de génération et la taille de la zone de gap. Ce traitement a été réalisé sur cinq stacks constitués de différentes électrodes supérieures. Nous avons ensuite pu comparer les résultats obtenus par notre modèle avec des simulations *ab initio*. Cela nous a permis de soulever des différences de fonctionnement, notamment avec une électrode à base de cuivre, qui présente un mécanisme hybride en CbRAM et OxRAM.

Maintenant que notre est modèle est opérationnel, nous allons tenter de le mettre à l'épreuve pour comprendre les principales failles existantes chez les mémoires résistives, à savoir leur forte variabilité, qui rend le travail sur la fiabilité plus compliqué qu'avec d'autres technologies de mémoires.

6. Références du chapitre III

[1] C. Yakopcic et al., “A Memristor Device Model”, IEEE Electron Device Letters, Vol. 32, No. 10, October 2011

[2] E. Miranda, D. Jiménez, and J. Suñé, “The Quantum Point-Contact Memristor ”, IEEE Electron Device Letters, Vol. 33, No. 10, October 2012

[3] J. J Yang, et al., "Memristive switching mechanism for metal/oxide/metal nanodevices", ” Nat. Nanotechnol., vol. 3, no. 7, pp. 429–433, July 2008

[4] D. Ielmini, “Modeling the universal Set/Reset characteristics of filament growth by field- and temperature- driven filament growth,” IEEE Transactions on Electron Devices, vol. 58, no. 12, pp. 4309–4317, December 2011

[5] S. Ambrogio, S. Balatti, DC. Gilmer and D. Ielmini, “Analytical modeling of Oxide-based bipolar Resistive Memories and complementary resistive switches IEEE Transactions on Electron Devices, Vol. 61, No. 7, July 2014

[6] A. Padovani, L. Larcher, G. Bersuker, and P. Pavan, “Charge transport and degradation in HfO2 and HfOx dielectrics,” IEEE Electron Device Letter, vol. 34, no. 5, pp. 680–682, May 2013

[7] https://www.scilab.org/fr

[8] H.-S.P. Wong, H.-Y. Lee, S. Yu et al., "Metal–Oxide RRAM," Proceedings IEEE, pp. 1951-1970, 2012

[9] X. Guan, S. Yu and H-S P Wong, “On the switching parameter variation of metal-oxide RRAM: I. Physical modeling and simulation methodology”, IEEE Transactions on Electron Devices, vol. 59, no. 4, April 2012

[10] B. Gao, B. Sun, H. Zhang, L. Liu, X. Liu, R. Han, J. Kang, and B. Yu, “Unified physical model of bipolar oxide-based resistive switching memory,” IEEE Electron Device Letters, vol. 30, no. 12, pp. 1326–1328, December 2009

[11] S. Yu, X. Guan and H-S P Wong, “Conduction mechanism of TiN/HfOx/Pt resistive switching memory: A trap-assisted-tunneling model”, Applied Physics Letters 99, 063507, 2011

[12] C. Walczyk et al., "Pulse-induced low-power resistive switching in HfO0 metal-insulator-metal diodes for nonvolatile memory applications," Journal of Applied Physics, vol. 105, no. 114103, 2009

[13] D. Ielmini, F. Nardi and C. Cagli, “Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories”, Nanotechnology 22, 254022, 2011

[14] Stefano Larentis et al., “Resistive Switching by Voltage-Driven Ion Migration in Bipolar RRAM—Part II: Modeling” IEEE Transactions on Electron Devices, vol. 59, no. 9, September 2012

[15] L. Larcher, A. Padovani, O. Pirrotta, L. Vandelli, and G. Bersuker, “Microscopic understanding and modeling of HfO2 RRAM device physics,” in Proc. IEEE Int. Electron Devices Meeting, Dec. 2012

[16] L. Larcher, " Statistical Simulation of Leakage Currents in MOS and Flash Memory Devices With a New Multiphonon Trap-Assisted Tunneling Model", Transactions on Electron Devices, Vol. 50, No. 5, May 2003

[17] D. Ielmini and Y. Zhang, " Analytical model for subthreshold conduction and threshold switching in chalcogenide-based memory devices ", Journal of Applied Physics 102, 054517, 2007

[18] D. Fugazza, D. Ielmini, S. Lavizzari, and A. L. Lacaita, "Distributed-Poole-Frenkel modeling of anomalous resistance scaling and fluctuations in phase-change memory (PCM) devices", IEEE International Electron Devices Meeting (IEDM) 2007

[19] D. Ielmini, " Threshold switching mechanism by high-field energy gain in the hopping transport of chalcogenide glasses", PHYSICAL REVIEW B 78, 035308, 2008

[20] C. Cagli, J. Buckley, V. Jousseaume et al., "Experimental and theoretical study of electrode effects in HfO2 based RRAM," IEEE Electron Devices Meeting IEDM, pp. 658–661, 2011

[21] R. B. Hall, "The Poole-Frenkel Effect", Thin Solid Films, 8 263-271, 1971

[22] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy and D. Ielmini, "Statistical Fluctuations in HfOx resistive-switching memory: Part II—Random Telegraph Noise", IEEE Transactions on Electron Devices, vol. 61, no. 8, august 2014

[23] Y. Li et al., "Conductance Quantization in Resistive Random Access Memory", Nanoscale Research Letters 10:420, 2015

[24] Y-E Syu et al., "Atomic-level quantized reaction of HfOx memristor", Applied Physics Letters 102, 172903, 2013

[25] E. Pérez et al., "Impact of the Incremental Programming Algorithm on the Filament Conduction in HfO2-Based RRAM Arrays", Journal of the Electron Devices Society, Vol. 5, No. 1, January 2017

[26] J. Suñé et al., "Electrical evidence of atomic-size effects in the conduction filament of RRAM", IEEE 11th International Conference on Solid-State and Integrated Circuit Technology (ICSICT), 2012

[27] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy and D. Ielmini, "Statistical Fluctuations in HfOx resistive-switching memory: part I – Set/Reset variability", IEEE Transactions on Electron Devices, vol. 61, no. 8, august 2014

[28] J.W. McPherson, J. Y. Kim, A. Shanware, H. Mogul, "Thermochemical description of dielectric breakdown in high dielectric constant materials", Appl. Phys. Lett., Vol 82(13), pp. 2121-2123, 2003

[29] J.W. McPherson and H. Mogul, "Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown in SiO_2 thin films", Journal of Applied Physics 84, 1513, 1998

[30] J. W. McPherson, R. B. Khamankar, and A. Shanware, "Complementary model for intrinsic time-dependent dielectric breakdown in SiO_2 dielectrics", Journal of Applied Physics 88, 5351, 2000

[31] A. Padovani et al., "Microscopic modeling of HfOx RRAM operations: from forming to switching", IEEE Transactions on Electron Devices, vol. 62, no. 6, June 2015

[32] X. Guan, S. Yu and H-S P Wong, "On the switching parameter variation of metal-oxide RRAM: I. Physical modeling and simulation methodology", IEEE Transactions on Electron Devices, vol. 59, no. 4, April 2012

[33] D. Berco and T-Y. Tseng, "A comprehensive study of bipolar operation in resistive switching memory devices", Journal of Computational Electronics, Volume 15, Issue 2, pp 577–585, June 2016

[34] K.S Ralls, D.C Ralph and R.A Buhrman, "Individual-defect electromigration in metal nanobridges", Physical Review B, vol. 40, no. 17, November 1989

[35] Z. Chen and R.S Sorbello, "Local Heating in mesoscopic systems", Physical Review B, vol. 47, no. 20, May 1993

[36] S. Blonkowski and T. Cabout, "Bipolar resistive switching from liquid helium to room temperature", Journal of Physics D : Applied Physics 48, 2015

[37] A. Padovani et al., “Understanding the Role of the Ti Metal Electrode on the Forming of HfO2-based RRAMs”, Memory Workshop (IMW), 4th IEEE International, 2012

[38] T. Cabout et al., “Role of Ti and Pt electrodes on resistance switching variability of HfO2-based Resistive Random Access Memory”, Thin Solid Films 533, 19–23, 2013

[39] www.mdlab-software.it

[40] B. Traoré et al., "HfO2-Based RRAM: Electrode Effects, Ti/HfO2 Interface, Charge Injection, and Oxygen (O) Defects Diffusion Through Experiment and Ab Initio Calculations", IEEE Transactions on Electron Devices, vol. 63. No. 1, January 2016

[41] S. Clima et al., "First-principles thermodynamics and defect kinetics guidelines for engineering a tailored RRAM device", Journal of Applied Physics 119, 2016

[42] C. Nail et al., "Understanding RRAM endurance, retention and window margin trade-off using experimental results and simulations", IEEE International Electron Devices Meeting, 2016

[43] Y. Bernard et al., “Back-end-of-line compatible Conductive Bridging RAM based on Cu and SiO2”, Microelectronic Engineering 88, 2011

[44] M.N. Kozicki, C. Gopalan, M. Balakrishnan, M. Park and M. Mitkova, “Non-Volatile Memory Based on Solid Electrolytes” in: Proceedings of the 2004 Non-Volatile Memory Technology Symposium (NVMTS), 2004

Chapitre IV : Etude du bruit et de son impact sur la fiabilité

1. Introduction

Ce quatrième chapitre du rapport de thèse abordera des thématiques reliées à la notion de fiabilité et de variabilité dans les mémoires résistives à base d'oxyde métallique. Pour explorer les problèmes de fiabilité, les tests sur des dispositifs unitaires, comme ceux que nous avons réalisés dans le deuxième chapitre de ce manuscrit, ne suffisent plus : il y a nécessité de passer à des tests sur des statistiques beaucoup plus importantes, qui vont alors nécessiter des mesures sur des matrices mémoires.

Dans une première partie de ce dernier chapitre nous présenterons donc un nouveau véhicule de test, conçu pour les tests matriciels. Nous expliquerons comment fonctionne le système d'adressage permettant l'adressage de matrice allant jusqu'au mégabit. Nous présenterons également le setup expérimental associé, qui diffère de façon importante du setup utilisé pour la caractérisation de cellules unitaires. Enfin nous expliquerons en quoi une telle combinaison véhicule de test / banc de test est nécessaire, et ce qu'elle nous apporte dans l'étude de la variabilité et de la statistique des mémoires OxRAM.

Ensuite nous confronterons notre modèle, qui n'a jusque-là servi qu'à modéliser le comportement de dispositifs unitaires, aux mesures réalisées sur les matrices. Notamment nous verrons dans quelle mesure notre modèle est capable de fitter les distributions de tension de switching en fonction de la vitesse des pulses utilisés, pour des matrices 4kb.

De même nous verrons l'influence du courant de forming utilisé sur l'évolution des distributions des tensions de switching. Le modèle nous permettra alors de fitter ces distributions en modulant le rayon du filament, ce qui nous laisse penser que le courant de forming commande le rayon de filament conducteur obtenu.

Enfin, nous nous consacrerons à l'étude du bruit RTN dans les mémoires OxRAM. En effet, de par la présence de défauts générés lors du forming et des différentes phases d'écriture et d'effacement, le courant peut être bruité en fonction de la présence, de la localisation ou de la charge électrique de ces défauts. Nous commencerons par décrire, en nous appuyant sur la littérature, les mécanismes physiques supposés à l'origine de ce bruit RTN dans les OxRAM et

en quoi il est intéressant de l'étudier, car c'est un paramètre qui va influencer la fiabilité des dispositifs (un dispositif très riche en bruit RTN sera moins fiable) et car son étude nous apporte des informations sur la structure microscopique du filament et sur la présence de défauts. Puis nous détaillerons notre protocole expérimental pour la mesure de ce bruit. Nous présenterons ensuite les résultats de ces mesures, avec notamment l'évolution à la fois du nombre de niveaux de bruit et de leur amplitude en fonction du courant de compliance utilisé lors du forming. Nous verrons ensuite quels sont les impacts directs de ces évolutions, tout d'abord sur des dispositifs unitaires, puis sur des matrices 4kb entières. Nous conclurons ensuite sur l'intérêt d'utiliser un forming à relatif fort courant : en effet, utiliser un forming faible résulte en une amplitude RTN plus forte qui a des répercussions sur les distributions HRS des matrices.

2. Présentation des tests sur matrices

Dans la seconde partie du second chapitre de ce rapport de thèse, nous avons souligné l'importante variabilité, caractéristique des mémoires OxRAM, notamment lorsque les temps de pulse, les courants de set ou les tensions de reset sont réduits, à cause du mécanisme filamentaire de switching [1, 2, 3, 4]. Lors des tests effectués sur la variabilité de nos cellules OxRAM, dans le second chapitre, nous avons majoritairement exploré la variabilité cycle-à-cycle : en effet nous avons mesuré 400 cycles sur 20 cellules. En effet, dans la mesure où nous n'avions à notre disposition que des dispositifs unitaires, il était compliqué d'étudier plus de quelques dizaines de cellules, et cela nécessitait un certain temps et était relativement contraignant pour l'opérateur.

Tous ces facteurs rendent l'utilisation de matrices de plus en plus indispensable pour mieux comprendre les problèmes de fiabilité de cette technologie. Avec des dispositifs sous forme de matrices et des bancs de test adaptés, il devient possible d'accéder à une statistique de mesure sans équivalent. Dans cette partie nous décrirons donc le véhicule de test adapté aux matrices OxRAM ainsi que le banc de test associé. Nous préciserons également le type d'information que nous pouvons extraire de ces mesures.

2.1. Structure du véhicule de test

Le véhicule de test dédié aux tests sur des matrices mémoires OxRAM (mais également CbRAM ou PCRAM) est baptisé MAD (Memory Advanced Demonstrator). Ce véhicule comprend à la fois des dispositifs unitaires et des matrices mémoires (de 256 bits à 1Mbits).

2.1.1/ Fabrication des dispositifs

De même que pour le véhicule MARS étudié dans le second chapitre, le dispositif mémoire est intégré en Back-End-Of-Line et les dispositifs sont en configuration 1T1R. Le processus de fabrication est donc divisé en deux parties. Les niveaux M1 à M4 sont fabriqués par STMicroelectronics en 130nm. Le LETI s'occupe ainsi du dispositif mémoire ainsi que du dernier niveau de métal, comme indiqué sur la Figure IV.1.

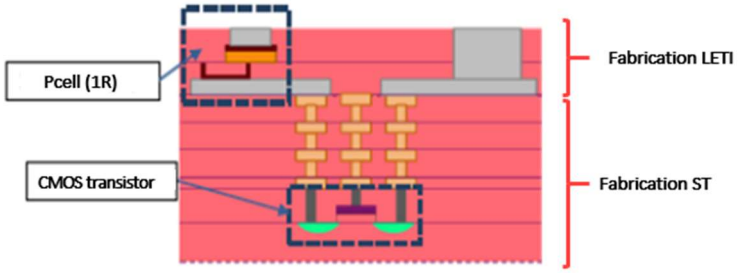


Figure IV.1 : Schématisation d'un dispositif 1T1R du véhicule MAD [5]

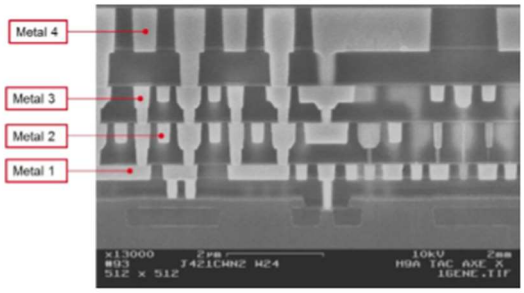


Figure IV.2 : Image TEM des niveaux M1 à M4 [5]

Les wafers MAD sont des wafers 200mm, contenant 52 puces. En Figure IV.3, nous avons représenté une puce telle qu'on en trouve sur nos wafers MAD. Même si un grand nombre de structures différentes existe sur une même puce, nous ne parlerons ici que des dispositifs mémoire 1T1R "classiques", sous forme de dispositifs unitaires ou de matrices.

De plus, comme nous l'avons dit plus haut, on peut signaler que ce véhicule est également adapté à d'autres technologies de mémoires émergentes non-volatiles, telles que les CbRAM ou les PCRAM. Seul le dispositif mémoire intégré en BEOL change. Cela permet une grande souplesse d'utilisation.

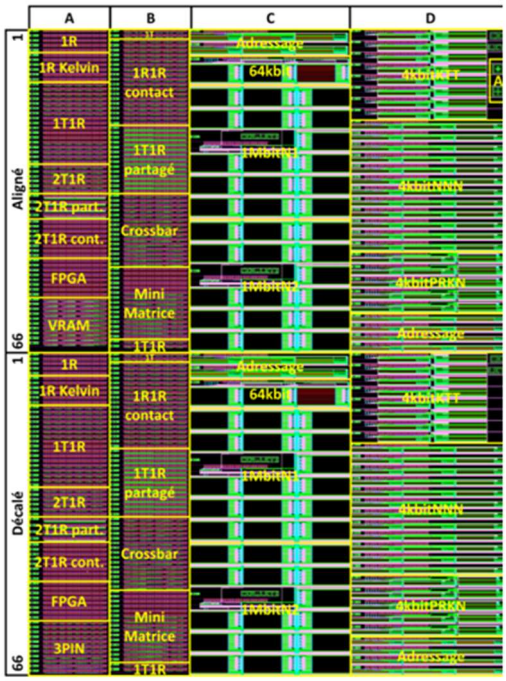


Figure IV.3 : layout d'une puce du véhicule MAD [5]

2.1.2/ Structure des matrices

Sur MAD, plusieurs tailles de matrices sont disponibles : 256bits, 4kbits, 65kbits et 1Mbits, toutes en configuration 1T1R. La bitline est connectée à l'électrode supérieure, la wordline à la grille du transistor tandis que la sourceline est connectée à l'électrode inférieure (cf. Fig. IV4).

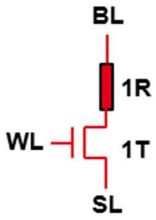


Figure IV.4 Schématisation d'un dispositif 1T1R et des connections associées.

Trois multiplexeurs assurent les fonctions de décodage pour les trois lignes. En fonction de la taille de la matrice, le nombre de lignes WL ou BL varient (la bitline et la sourceline partagent la même adresse). Ainsi :

- les matrices 256bits sont composées d'une WL et de 256 BL, adressées par 8 bits d'adressage.
- les matrices 4kbits sont composées de 16 WL, adressées par 4 bits d'adressage, et de 256 BL, adressées par 8 bits d'adressage.
- les matrices 65kbits sont composées de 256 WL, adressées par 8 bits d'adressage, et de 256 BL, adressées également par 8 bits d'adressage.

- les matrices 1Mbits sont composées de 16 matrices 65kbits, appelées secteurs. Le secteur est sélectionné via 4 bits d'adressage (appelées WLm), tandis que l'adressage au sein d'un secteur se fait comme pour une matrice 65kbits classiques, c'est-à-dire avec des MUXs de 8 bits pour les WL et BL.

Le schéma représenté en Figure IV.5 explique la structure d'une matrice 1Mbits. L'image est tirée de [5].

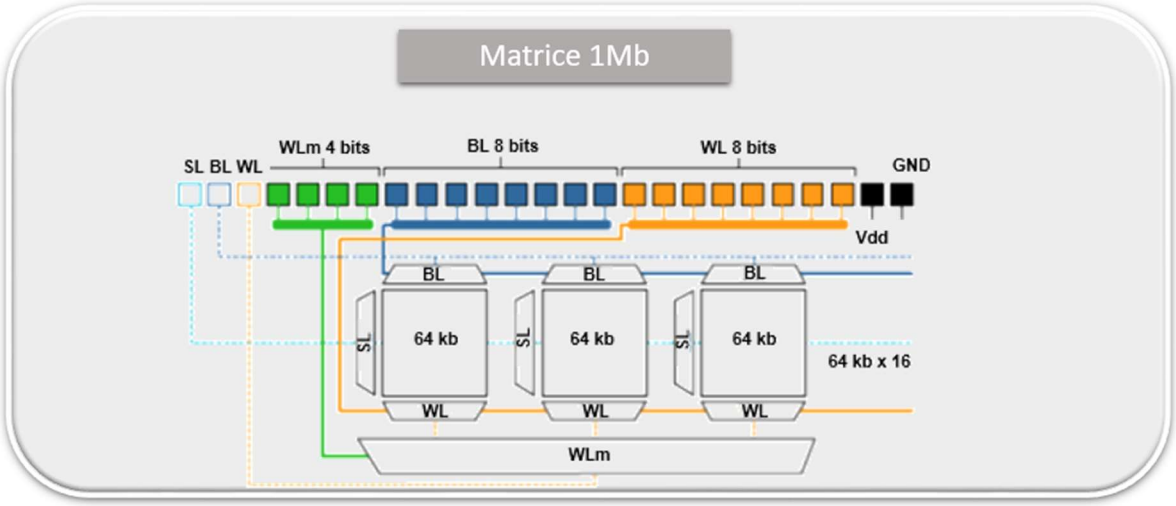


Figure IV.5 : Structure d'une matrice 1Mbits [5].

Pour programmer ou lire une matrice d'un mégabit, 25 pointes sont nécessaires :

- 3 pointes analogiques, avec les signaux électriques de set/reset/read
- 4 pointes pour l'adressage du secteur
- 8 pour celui de la BL
- 8 pour celle de la WL
- 1 pour l'alimentation
- 1 pour la masse

Ainsi, tous les plots du scribe sont utilisés. Au contraire, pour le cas d'une matrice 4kbits, seuls 17 plots sont nécessaires :

- les 3 pointes analogiques
- 8 pointes pour la BL
- 4 pointes pour la WL
- Les deux pointes d'alimentation et de masse

Ainsi, trois matrices sont présentes par scribe (cf. Figure IV.6).

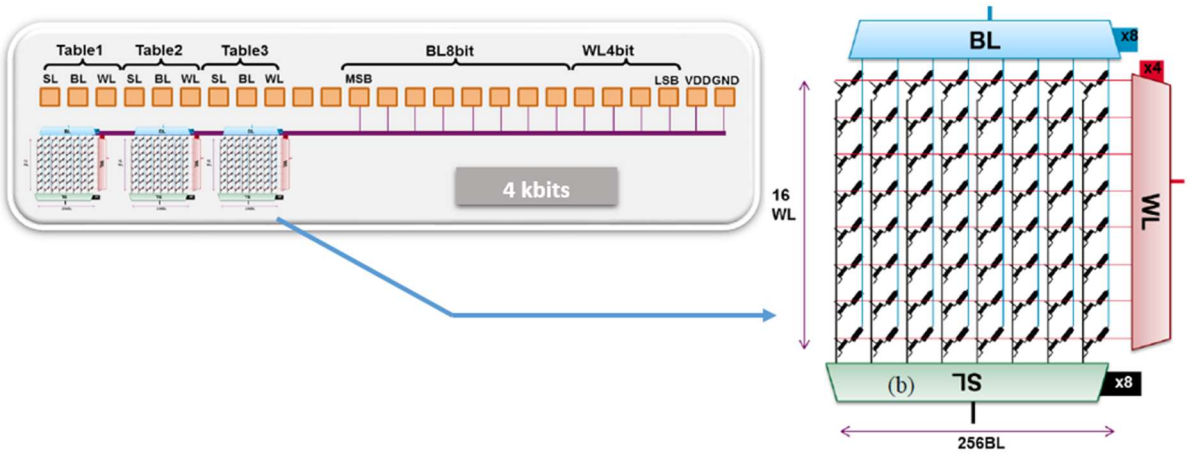


Figure IV.6 : Structure d'un scribe 4kbits et des matrices correspondantes [5].

En figure IV.7, nous avons représenté une image SEM d'une telle matrice 4kbits. Le dispositif mémoire est entouré en blanc.

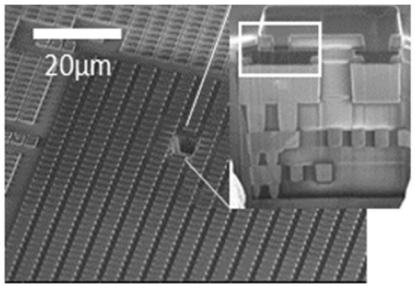


Figure IV.7 : Image SEM d'une matrice 4kbits

2.2. Présentation du banc de test pour les tests sur matrices

Maintenant que nous avons présenté le réticule MAD dédié aux tests sur des matrices de mémoires résistives, nous allons présenter le banc de caractérisation électrique associé. Celui-ci a été spécifiquement dédié au véhicule MAD.

Le set-up peut être divisé en deux composantes : une composante analogique, responsable des opérations de set, reset et read sur les mémoires, et une composante digitale, responsable de l'adressage des multiplexeurs. La partie analogique est exercée via un PGU (Pulse Generator Unit) Keysight B1530 [6]. Celui-ci s'occupe d'envoyer des pulses de tensions et de mesurer le courant sur les BL, SL ou WL. La résolution en courant, sur le plus petit calibre en courant, c'est-à-dire 1nA, est de 1fA et le taux d'échantillonnage est de $0.2ns^{-1}$. Cette partie analogique est assez similaire à celle employée dans le second chapitre de ce manuscrit de thèse.

La partie digitale est assurée par un microcontrôleur Arduino ATmega2560 [7] (cf. Fig. IV.8). Il permet d'envoyer les adresses aux mémoires à écrire/effacer/lire aux différents décodeurs. Celui-ci est connecté à la sortie Trigger du B1500. Lorsque ce dernier a fini sa

mesure, il envoie un signal de trigger au microcontrôleur, qui s'occupe alors de changer l'adresse.

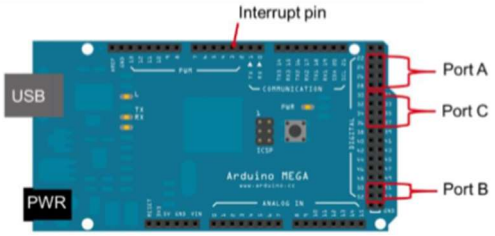


Figure IV.8 : microcontrôleur Arduino [5].

Le schéma de fonctionnement global des mesures de caractérisation électrique est représenté en Figure IV.9.

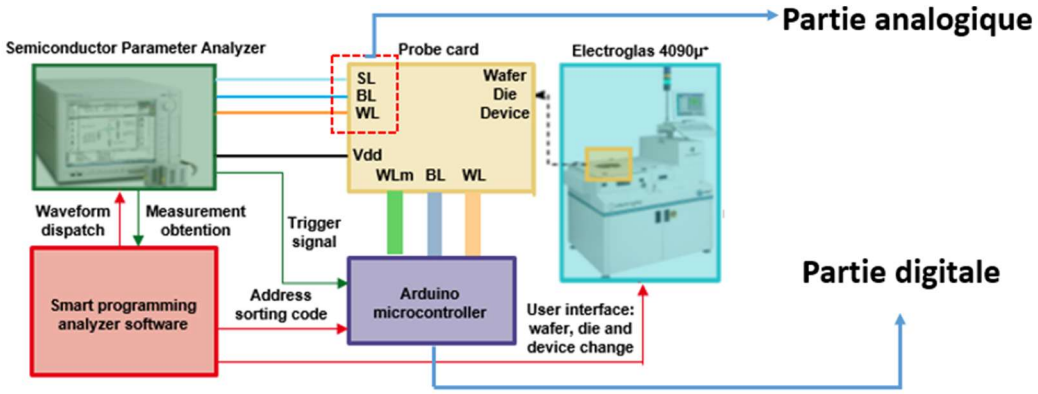


Figure IV.9 : Set-up expérimental du test sur matrices.

Comme on peut le voir sur ce schéma, une carte de mesure est utilisée pour connecter les 25 plots aux pointes (cf. Figure IV.10). Les plaques à étudier sont placées sur un banc de test Electroglas 4090µ^{+} [8]. 25 câbles SMA sont utilisés pour connecter la carte soit au microcontrôleur soit aux sorties du PIV (cf. Figure IV.11). Enfin, la tension d'alimentation V_{dd} de 4.8V est soit fournie par le microcontrôleur soit par une alimentation externe.

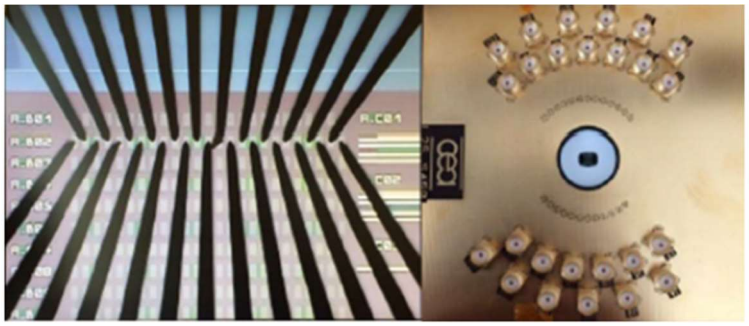


Figure IV.10 : image de la carte de mesure et des 25 pointes connectées au DUT [5].

Un point important est qu'il n'y a pas de communication entre le B1530 et le microcontrôleur. Lorsque le B1530 a fini sa mesure, il envoie un signal de trigger au microcontrôleur qui change ensuite l'adresse. Cependant, l'Arduino ne renvoie pas de signal au B1530 lorsque le changement d'adresse a été correctement effectué (cela serait trop couteux en

temps). Si le timing n'est pas bon, le B1530 pourrait lancer la nouvelle série de mesure avant que le changement d'adresse n'ait été effectué. La mesure du temps entre l'envoi du signal de trigger du B1530 vers l'Arduino et les pulses de changement d'adresse envoyés par celui-ci vers les multiplexeurs est donc importante à prendre en compte.

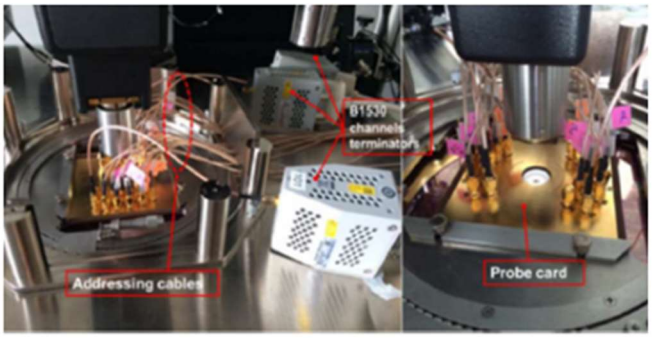


Figure IV.11 : image des connections et de la carte de mesure [5].

Pour mesurer ce délai, travail réalisé dans [5] par L. Kadura, la sortie Trigger du B1500/B1530 a été connectée sur la voie 1 d'un oscilloscope. Les 3 autres voies sont connectées aux sorties LSB (Lower Significant Bit) de l'Arduino. Pour ces tests, ces trois ports sont programmés pour renvoyer 5V après réception du signal de trigger. Le montage expérimental est représenté en Figure IV.12.

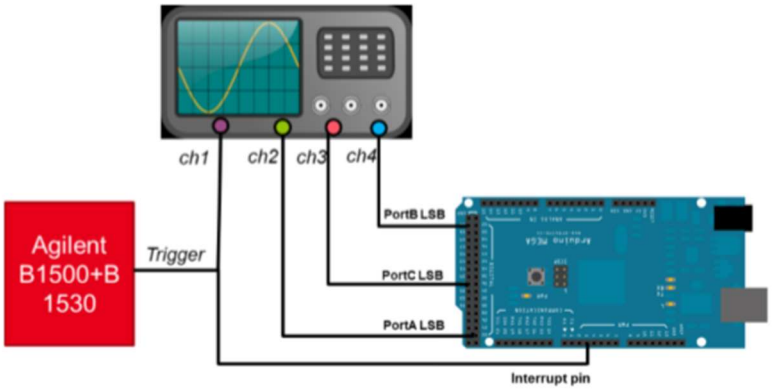


Figure IV.12 : Schéma descriptif de la mesure du délai [5].

En Figure IV.13, une capture d'écran d'une telle mesure est représentée. On constate bien un délai, d'environ 4µs ici entre l'envoi du signal de trigger et les LSB des trois ports (chacun espacé d'environ 180ns).

Un temps d'exécution de l'ISR (Interrupt Service Routine), qui représente le temps de calcul de la nouvelle adresse doit également être pris en compte. En Tableau 1, les mesures de ces deux temps sont résumées. Afin de s'assurer d'un bon timing entre le microcontrôleur et le B1530, un temps d'attente de 20µs entre le pulse de trigger et le début de la série de mesure a été implémenté.

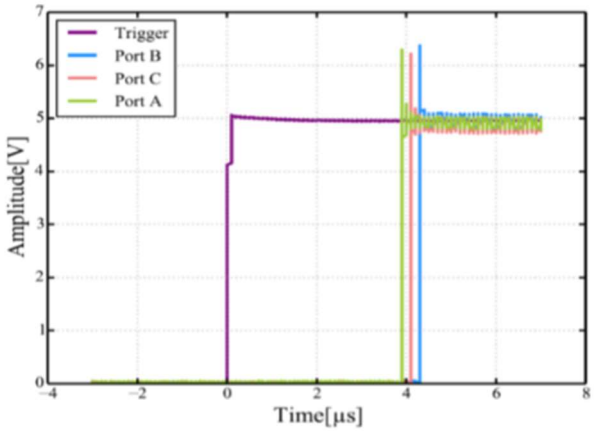


Figure IV.13 : Mesure du délai à l'oscilloscope [5].

	Trigger-Port	ISR
Mean	4.37 µs	1.93236 µs
Min	4.23 µs	1.56 µs
Max	**10.74 µs**	**6.46 µs**
sDev	300.2 ns	93.99 ns
No.of samples	70.574e+3	163.840e+3

Tableau IV.1 : mesure des différents délais [5]

2.3. Présentation des tests sur les mémoires

L'intérêt principal d'une telle configuration véhicule de test / banc de mesure est l'accès à une statistique bien supérieure à celle obtenue sur un banc de test classique, notamment en ce qui concerne la variabilité cellule-à-cellule.

2.3.1/ Coût temporel des opérations sur matrices

Dans la mesure où le banc de test doit permettre de mesurer des matrices allant jusqu'au Mbit, sur un très grand nombre de cycle, il est important de s'assurer que la durée des tests ne soit pas excessivement élevée. Nous nous intéresserons à analyser le temps de mesure nécessaire à la caractérisation d'une matrice d'un mégabit. Pour ce test, supposons que nous désirions réaliser 100 cycles d'écrite/lecture/effacement/lecture sur chacune des cellules de la matrice, il y a donc trois différentes constituantes temporelles : le temps d'écriture/effacement ($1M*(t_{SET}+t_{RESET})*100$), le temps de lecture ($1M*2*t_{READ}*100$) et le temps d'adressage ($1M*20\mu s=20s$). Si on prend le test décrit en Figure IV.14.a, on obtient la répartition temporelle décrite en (b). On se rend facilement compte que le temps d'adressage n'est clairement pas le facteur limitant. L'opération de lecture est très nettement l'opération la plus coûteuse en temps. Il s'agit d'une bonne nouvelle : alors qu'on pouvait penser que le temps d'adressage (non compressible) pourrait être un obstacle, ce n'est pas le cas. L'opération de read, quant à elle peut

être facilement réduite, puisque les conditions utilisées ici étaient très relaxées par rapport aux performances dont les OxRAM sont capables, en termes de temps de lecture, de set et de reset.

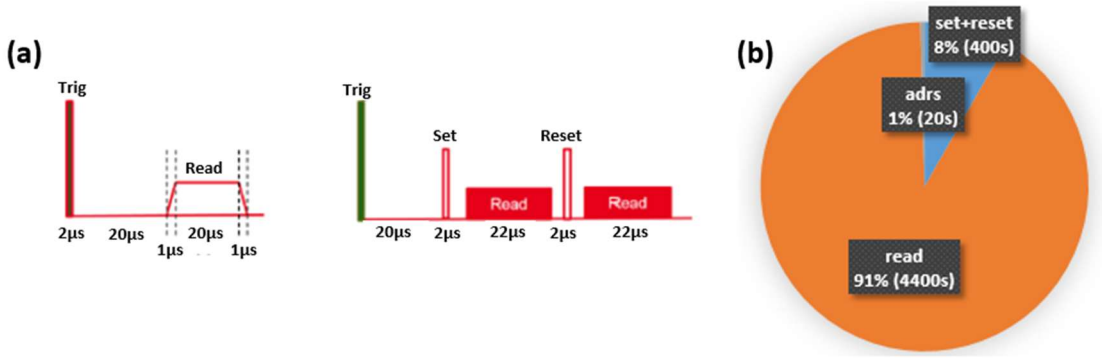


Figure IV.14 : (a) schéma des pulses de reset, set et reset [5]. (b) Répartition du cout temporel des différentes opérations.

En plus des tests très classiques de cyclages, tels que ceux décrits ci-dessus, plusieurs programmes sur ce banc de tests permettent de réaliser un grand nombre de tests différents, dont les principaux sont représentés en Figure IV.15.

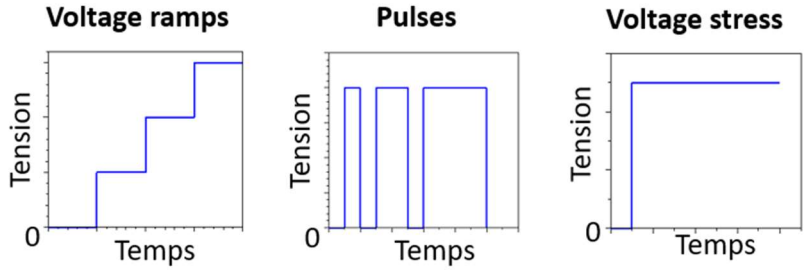


Figure IV.15 : Exemple de différents types de signaux.

Ces différents modes de programmations permettent d'accéder à différents types de données. Dans les tests qui suivront nous utiliserons principalement le premier (rampe de tension en escalier) et le dernier (tension constante).

2.3.2/ Test d'homogénéité des matrices

L'un des autres grands intérêts d'avoir accès à des tests automatisés sur un grand nombre de dispositifs est que l'on peut contrôler leur homogénéité spatiale : cela consiste à regarder si, le long d'une ligne (BL par exemple), il y a une déviation de la résistance. Le véhicule de test MAD contient des matrices sans dispositifs mémoires, uniquement dotées d'un transistor. On peut commencer par analyser les déviations spatiales de résistance sur une telle matrice, par exemple, ici de 4kbits.

L'opération consiste à réaliser une bitmap de la matrice à étudier. Il s'agit simplement d'une mesure de la résistance de chaque bit de la matrice. La bitmap est la représentation spatiale de la distribution de résistance en fonction du numéro de BL et de WL. La bitmap d'une matrice 4 kbit sans dispositifs mémoire est représentée en Figure IV.16. On constate qu'il y a une légère déviation de résistance le long de la bitline : de la première à la dernière, la résistance augmente d'environ 15Ω, soit une augmentation de 3%. Ceci s'explique tout simplement par la plus grande longueur de bitline pour les dispositifs en bout de ligne. L'influence de cette déviation ne devrait pas avoir d'impact sur le fonctionnement de la matrice. On remarque qu'il n'y a pas de déviation le long de la wordline. En effet celle-ci est connectée à la grille des transistors et aucun courant n'y circule.

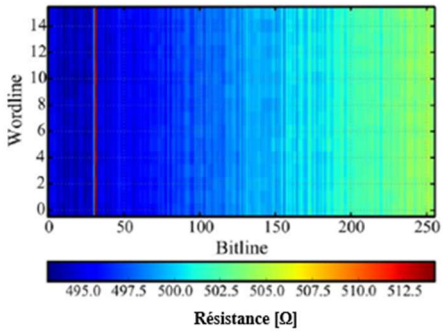


Figure IV.16 : bitmap d'une matrice 4kbit, sans dispositifs mémoires. La BL n°31 a été volontairement déconnectée, d'où la forte résistance, bien observée grâce à la ligne rouge correspondante. [5]

On peut également réaliser des bitmaps sur des matrices plus importantes, et dotées d'éléments mémoires. Prenons ainsi l'exemple d'une matrice 1Mbits OxRAM. Celle-ci, après forming a subi un simple cycle d'écriture/lecture/effacement/lecture. Pour les deux lectures on peut représenter, via un code couleur, la résistance de chaque mémoire en fonction de leur emplacement. La bitmap obtenue est représentée en Figure IV.17. Celle-ci est découpée en 16 parties, qui sont les 16 secteurs de 65kbits constituant les matrices 1Mbits. On ne remarque ici aucune déviation spatiale de résistance : la matrice est bien uniforme. En effet les fluctuations de résistances apparaissent de façon aléatoire. On peut également conclure que la matrice est fonctionnelle : on écrit et efface correctement les dispositifs mémoires.

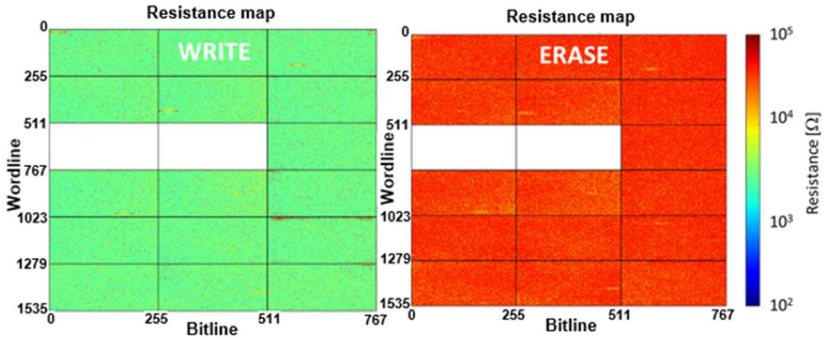


Figure IV.17 : bitmap d'une matrice 1Mbits.

2.3.3/ Test de variabilité

Le gros atout d'un tel set-up est bien sûr l'accès à une étude de la variabilité cellule-à-cellule bien plus poussée. Dans cette partie nous ne parlerons que de variabilité cellule-à-cellule, puisqu'il s'agit de la forme de variabilité que nous avons jusque-là le moins étudiée et qui est rendue possible avec ce set-up. En effet, sur les courbes de distributions normales à échelle de déviation standard, l'exploration des zones au-delà de 3σ (c'est-à-dire permettant de percevoir les 0.3% les plus éloignés de la moyenne) nécessite l'étude de plusieurs milliers de dispositifs, ce qui n'est, en pratique, possible qu'avec l'utilisation de matrice kilobits, ou plus. Il est très important d'avoir accès à cette zone de 3σ et au-delà. Comme cela est illustré sur la Figure IV.18, on observe très souvent une distribution de résistance gaussienne, sur les zone 1σ et 2σ. Ici nous avons représenté les résistances ON (après écriture) et OFF (après effacement) d'une matrice 4kbits OxRAM.

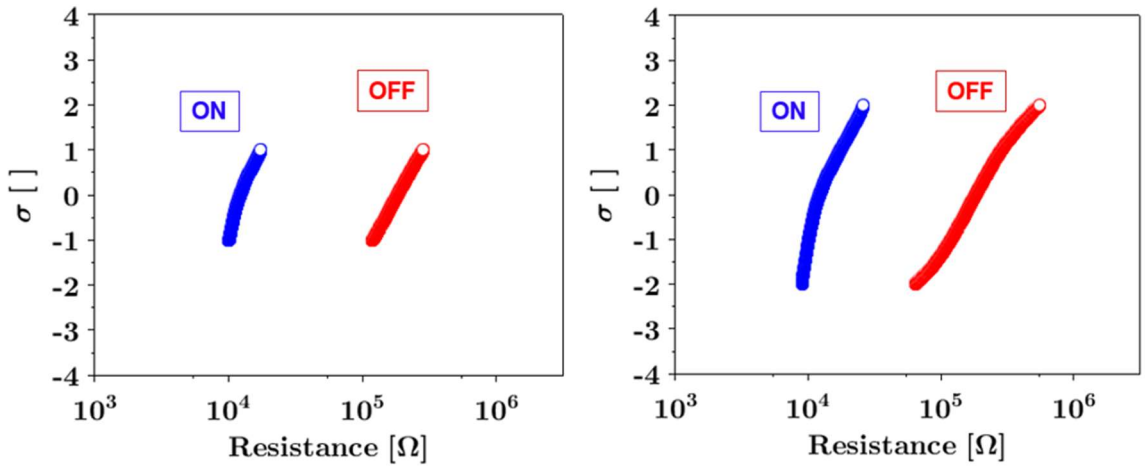


Figure IV.18 : Distributions sur une matrice 4kbits sur les zones 1 et 2σ.

L'accès à des matrices kilobits permet d'aller sonder au-delà de 3σ, et d'obtenir la courbe représentée en Figure IV.19.

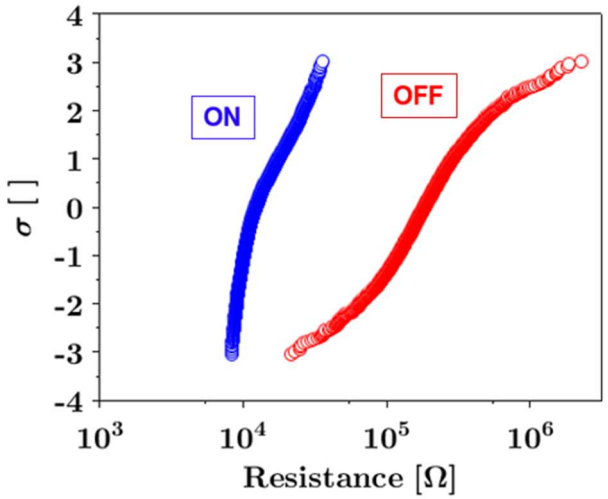


Figure IV.19 : Distribution sur la même matrice 4kbit au-delà de 3σ.

Au-delà de 2σ, des queues de distributions commencent à apparaître, notamment sur les distributions HRS. Comme cette zone ne concerne que quelques dizaines de pourcents des cellules, sans matrice kilobits, on n'y est que très peu sensible. Sur cette courbe on remarque également que la courbe LRS n'est pas gaussienne, et qu'elle semble composée de deux contributions : la contribution du transistor à gauche et la contribution intrinsèque des OxRAM à droite.

On constate ici que les distributions LRS et HRS, pourtant assez distinctes jusqu'à 2σ, interfèrent et qu'il n'y a pas ici de fenêtre résistive entre les dispositifs les plus résistants de la courbe LRS et les moins résistants de la courbe. L'utilisation de matrices kilobits est donc obligatoire pour attester de la fiabilité cellule-à-cellule des OxRAM.

2.3.4/ Mesures de tension de switching

Pour recentrer le sujet sur le domaine des tensions de switching, ce nouveau set-up nous fournit une nouvelle voie pour analyser les tensions de set des mémoires, de façon automatisée. Pour ceci, nous emploierons une tension d'électrode supérieure en RVS (Ramp Voltage Stress, cf. Fig. IV.20). Lors de ces mesures, le courant est mesuré toutes les 40µs, à chaque pas de tension. Les dispositifs étudiés ici sont 500 dispositifs OxRAM, issus de matrices 4kbits, de TiN/Ti/$HfO_2$5nm/TiN, à la fois en écriture et en effacement. Le but est de regarder l'influence de la vitesse de rampe sur les tensions de set et de reset, mais sur des statistiques bien supérieures aux mesures présentées dans le second chapitre de ce manuscrit. Nous avons donc comparé deux vitesses de rampes : 32 et 1600V.s^{-1} et mesuré les distributions de tensions de set et de reset. Les courbes de distributions pour set et reset, pour les deux vitesses sont représentées en Figure IV.21.

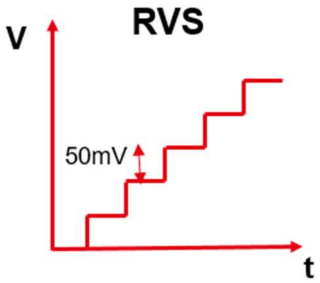


Figure IV.20 : schéma des pulses de tensions envoyés.

Que ce soit pour les opérations de set et de reset, comme nous l'avons vu dans le second chapitre du manuscrit, l'emploi d'une vitesse de rampe élevée augmente les valeurs de tensions de switching. On constate également, et ceci est une information nouvelle, que l'augmentation de la vitesse de rampe augmente aussi la dispersion (ceci est particulièrement visible sur le reset).

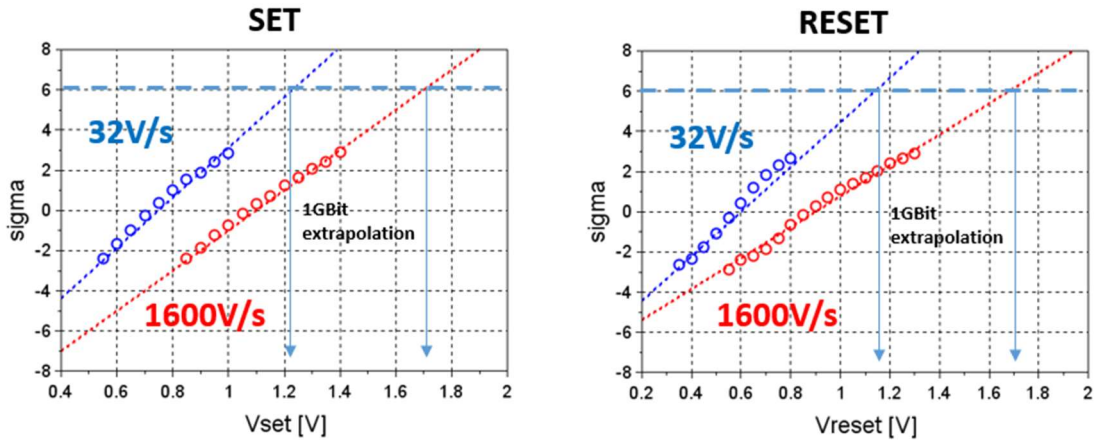


Figure IV.21 : Distribution des tensions de set et de reset pour les deux vitesses de rampe, et extrapolation pour des matrices Gbits.

Les distributions semblent approximativement gaussiennes. Pour le set, nous obtenons une valeur moyenne de 0.75V et un écart type de 0.08 à une vitesse de 32V.s^{-1}, et une valeur moyenne de 1.1V et un écart type de 0.10V à une vitesse de 1600V.s^{-1}. Pour le reset, nous obtenons une valeur moyenne de 0.6V et un écart type de 0.09 à une vitesse de 32V.s^{-1}, et une valeur moyenne de 0.9V et un écart type de 0.13V à une vitesse de 1600V.s^{-1}. Cela rend possible l'extrapolation, en prolongeant les droites de distributions, pour des tailles de matrices plus importantes. Ici nous avons extrapolé pour des matrices 1Gbits (à 6σ) : c'est-à-dire que nous avons extrapolé les valeurs minimales de tension nécessaires aux écritures et effacements de matrices de 1Gbits entières. Ainsi pour écrire une matrice 1Gbit, il faut employer une tension minimale de 1.22V à une vitesse de 32V.s^{-1} et de 1.71V à une vitesse de 1600V.s^{-1}. De même, pour effacer une matrice 1Gbit, il faut employer une tension minimale de 1.16V à une vitesse de 32V.s^{-1} et de 1.71V à une vitesse de 1600V.s^{-1}. Pour s'assurer d'une fenêtre résistive convenable il est néanmoins préférable d'utiliser des tensions plus élevées.

Ces travaux sur les matrices OxRAM du véhicule MAD ont fait l'objet d'une présentation orale, à l'ICMTS 2017(IEEE International Conference on Microelectronic Test Structures), à Grenoble [9].

3. Fit des distributions 4kbits des tensions de switching par le modèle pour différents temps de pulse

Le but de cette partie est de voir dans quelle mesure le modèle que nous avons présenté dans le troisième chapitre de ce manuscrit peut s'adapter aux mesures à plus grandes statistiques, permises par ce nouveau set-up.

Pour cela, une manipulation similaire à celle présentée ci-dessus, est réalisée. Elle consiste, une nouvelle fois, à faire varier la vitesse de rampe, et à chaque pas en tension (de 0.1V), à mesurer le courant, pour déterminer quelles sont les cellules qui ont switché, avant d'entamer le pas suivant. Nous avons donc pris deux matrices 4kbits : une matrice "classique" TiN/Ti10nm/$HfO_2$10nm/TiN et une matrice "hybride" TiN/$CuTe_2X$10nm/$HfO_2$10nm/TiN. Après forming, ces deux matrices ont été effacées puis nous avons réalisé la mesure de tension de set des matrices entières, pour trois temps de montée différents (10µs, 1ms et 10ms). On obtient alors les courbes de distributions des tensions de set.

Pour fitter ces distributions avec le modèle, il nous fallait introduire un paramètre aléatoire : en effet, notre modèle est purement déterministe. Seuls les paramètres d'entrée conditionnent le résultat de la simulation. Nous avons donc supposé une distribution gaussienne de g (la longueur de la zone de gap). Pour les dispositifs classiques, les fits ont été obtenus en centrant cette distribution sur une valeur moyenne de 2.4nm avec un écart type de 20% (soit 0.625nm). Pour les dispositifs hybrides, cette distribution est centrée sur une valeur moyenne de 1.5nm et un écart type de 10% (cf. Fig. IV.22). Pour que la simulation ne prenne pas trop de temps, nous n'avons répété les simulations que 200 fois (au lieu de 4000 fois, pour 4kbits).

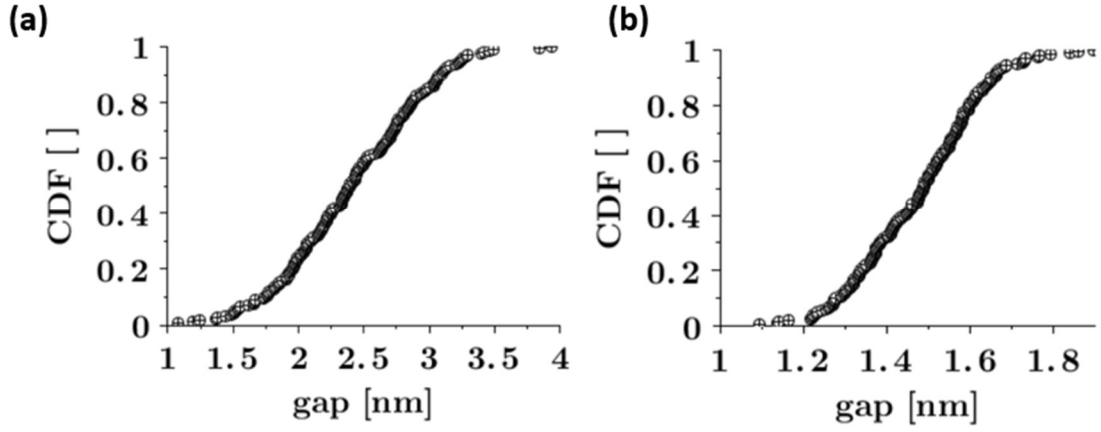


Figure IV.22 : Distribution normale des gaps pour les dispositifs (a) avec électrode de Ti et (b) avec électrodes de $CuTe_2X$.

Le fait que nous ayons dû prendre un gap moins dispersé (10% contre 20%) pour la matrice avec électrode à base de cuivre laisse penser que la variabilité sur le gap, avec cette électrode est inférieure à celle obtenue avec une électrode classique de titane.

La comparaison entre les résultats expérimentaux et les simulations est représentée, pour les deux stacks, en Figure IV.23.

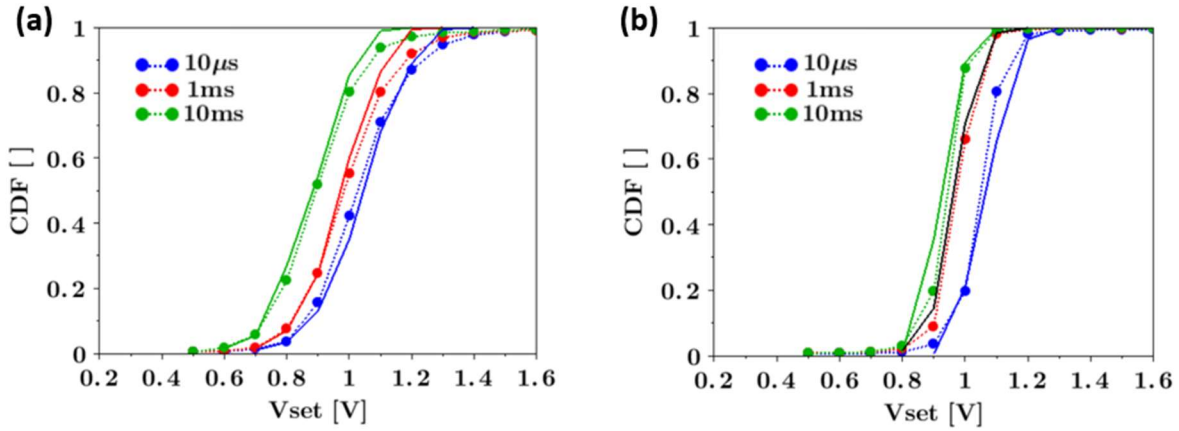


Figure IV.23 : Mesures expérimentales et simulations des distributions (simulations en traits pleins) pour des dispositifs avec (a) électrodes en titane et (b) électrodes en $CuTe_2X$.

Les simulations décrivent bien le gros des distributions et l'évolution de la valeur moyenne des distributions. Cela confirme que la variabilité peut être attribuée, en partie du moins, à une irrégularité de la valeur du gap g, d'un dispositif mémoire à un autre. En effet pour toutes ces simulations, les énergies d'activations ont été laissées constantes (2.75eV pour Ti/HfO_2 et 3.7eV pour $CuTe_2X$, soit des valeurs très proches de celles déterminées dans le chapitre précédent), de même que le rayon du filament qui a toujours été laissé à 1nm.

En revanche, principalement pour les dispositifs à électrode supérieure en titane, les queues de distributions ne sont pas correctement modélisées. Cela peut être liée au fait que la taille du gap n'est en réalité pas réellement gaussienne.

Ces manipulations révèlent l'une des raisons pour lesquelles nous avons opté pour un modèle semi-analytique plutôt simple : il nous permet de lancer, pour des durées raisonnables, quelques centaines de simulations (ici 200 pour chaque courbe), ce qui n'est pas possible avec des modèles numériques plus élaborés. En conservant une volonté de reproduire la physique des OxRAM, nous parvenons à reproduire également des résultats statistiques.

4. Impact du forming : fit avec le rayon du filament

Ainsi équipés d'un set-up expérimental capable de traiter des matrices de plusieurs milliers de dispositifs OxRAM, et d'un modèle capable de simuler le comportement d'un grand nombre de dispositifs sur un temps raisonnable, nous avons voulu analyser l'impact du courant de forming sur les distributions de tension de set, après cyclage. En effet, la question de savoir si le courant de forming a un impact sur la façon dont se comporte les cellules, même après

cyclage, est intéressante, et pourrait nous apporter à la fois des informations sur la structure microscopique du filament, mais également des clés pour améliorer la fiabilité des mémoires OxRAM.

Nous avons encore décidé de travailler sur les matrices 4kbits du véhicule MAD, sur les dispositifs "classiques" de TiN/Ti10nm/$HfO_2$10nm/TiN. Le principe de ces mesures est expliqué en Figure IV.24 : trois matrices 4kbits sont formées à des courants de forming différents (50µA, 100µA et 150µA), avec des pulses à 4V de 10µs. Ces 3 matrices subissent ensuite un léger cyclage de 100 opérations de set/read/reset/read, avec des pulses de set et de reset respectivement de 2 et 2.5V longs de 1µs, et un courant de compliance de 50µA pendant le set, pour les trois matrices. Ainsi, seule l'opération de forming diffère, entre les 3 matrices. Enfin, une tension en RVS, à une vitesse de 100µs, est appliquée sur la top électrode, avec des lectures à chaque pas de tension (de 0.1V), comme pour la manipulation présentée en partie 3 de ce même chapitre, pour déterminer la tension de set pour chaque dispositif.

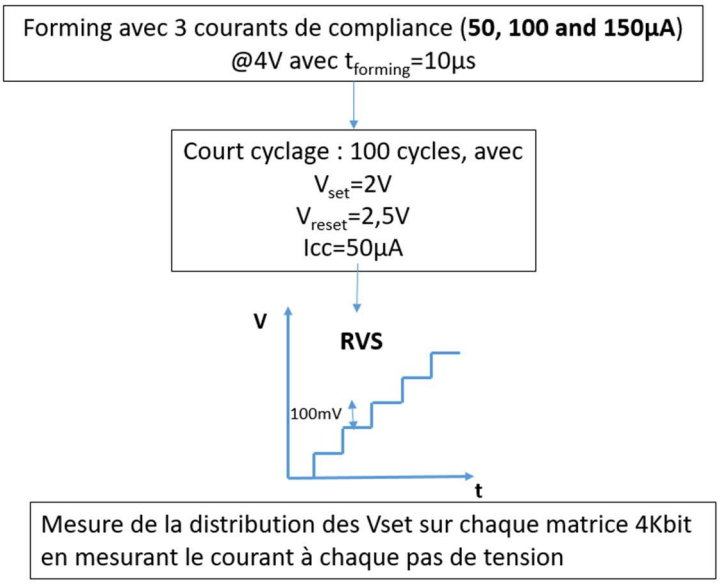


Figure IV.24 : Schéma explicatif de la manipulation réalisée pour tester l'influence du courant de forming sur les tensions d'écriture.

On peut alors tracer, pour les trois matrices, les courbes de distributions, à échelle de déviation standard, en fonction de la tension d'écriture. Ces courbes sont représentées en Figure IV.25, avec un zoom sur les valeurs proches de la moyenne de ces trois distributions. A noter que la partie inférieure à 0.5V est grisée car les lectures n'étaient commencées qu'à partir de cette tension de 0.5V.

Même après un cyclage de 100 cycles, on constate sans difficulté que l'intensité du courant de forming employé continue à affecter les cellules. En effet, plus le courant de forming

utilisé était intense, plus la tension à laquelle la cellule est écrite diminue. Ceci est plus facilement visible sur la courbe en inset, où le décalage des valeurs moyennes de tension de set est net : la tension de set diminue d'environ 100mV entre 50 et 150µA.

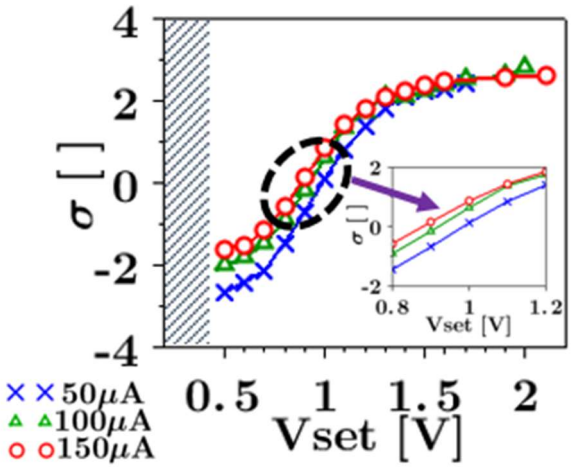


Figure IV.25 : Impact du courant de forming sur les tensions de set des matrices, après un cyclage de 100 cycles d'écriture/effacement des cellules.

On voit clairement que la distribution n'est ici pas gaussienne (sinon, les distributions seraient des droites). C'est pourquoi, nous ne tenterons pas, avec notre modèle de modéliser ces distributions, mais uniquement les valeurs moyennes de celles-ci.

Beaucoup de modèles relient le courant de compliance à l'épaisseur du filament conducteur, ou à la densité de lacunes d'oxygène [10, 11, 12, 13, 14]. Plus le courant de compliance utilisé est important plus on est supposé générer de lacunes d'oxygène et ainsi obtenir un filament plus épais. Nous avons donc tenté de vérifier si le fait de varier le rayon du filament, lors de nos simulations permettait de suivre l'évolution de la tension d'écriture des mémoires. Ainsi, nous avons réalisé des simulations en laissant le gap et l'énergie d'activation de génération de lacunes d'oxygène constants (respectivement à 2.3nm et à 2.7eV). Nous avons ensuite fait uniquement varier le rayon du filament conducteur pour fitter les valeurs moyennes des courbes de la Figure IV.25. Nous sommes parvenus à bien reproduire l'évolution de Vset en fonction du courant de forming en utilisant des rayons de filaments de 1nm, 2nm et 4nm pour les courants de 50µA, 100µA et 150µA respectivement. La courbe obtenue est représentée en Figure IV.26.

Faire varier le rayon du filament permet de reproduire très efficacement la courbe d'évolution de la valeur moyenne des distributions de tension de set. Augmenter le rayon du filament augmente le nombre de lacunes qui peuvent être générées et rend donc l'opération de set plus facile. Cela indique donc que le forming a majoritairement une influence sur le rayon du filament et que le filament "garde en mémoire" l'intensité du courant de forming, puisque

malgré un cyclage de set/reset de cent cycles identiques, le rayon du filament reste fixé par le courant de forming. Cependant cela reste sous réserve que le courant de set utilisé durant le cyclage n'excède pas celui utilisé durant le forming. En effet, ici, tous les cyclages ont été effectués à 50µA. Ce que l'on peut conclure, c'est que les dispositifs ayant vu un courant de forming supérieurs (100 ou 150µA) restent marqués par ce courant plus intense, via un rayon de filament plus grand. En Figure IV.27, nous avons ajouté un schéma décrivant la différence, entre un forming à fort et à faible courant, sur les opérations de set et reset. Ces deux opérations agissent sur la zone de gap, mais pas sur le rayon du filament qui reste fixé par le forming.

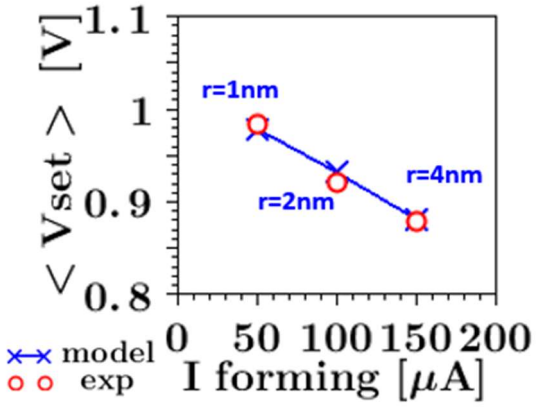


Figure IV.26 : Evolution de la tension moyenne de set en fonction du courant de forming (points expérimentaux), fittée par notre modèle par une évolution du rayon du filament (de 1nm, 2nm et 4nm pour les courants de 50µA, 100µA et 150µA).

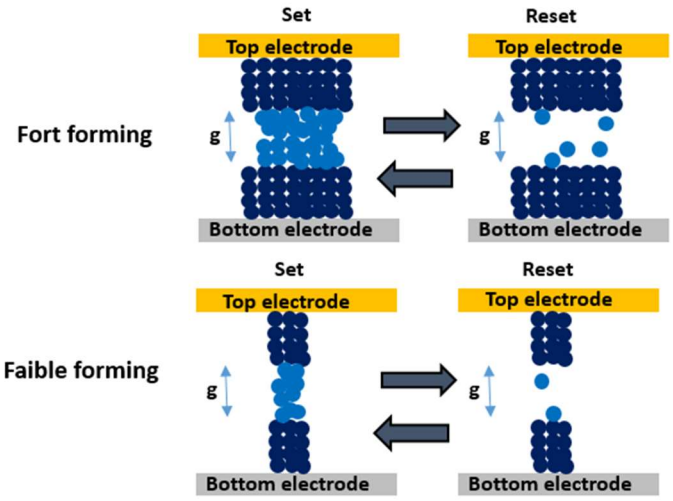


Figure IV.27 : Illustration de la différence entre fort et faible forming sur le filament conducteur.

Nous avons donc démontré que le courant employé durant le forming continue à influencer le comportement des mémoires, sur le long terme et que cela pouvait être lié à la variation du rayon du filament conducteur. Nous voulons désormais voir si cela a un impact sur la fiabilité des dispositifs, et notamment sur les mesures de bruits à basse fréquence (bruit RTN : Random Telegraph Noise principalement).

5. Mesure de bruit à basse fréquence

Dans cette partie, nous présenterons les mesures de bruit à basse fréquence que nous avons réalisés sur nos dispositifs OxRAM. Avant cela, il convient au préalable de présenter dans une première sous-partie, ce qu'est ce bruit à faible fréquence et quelles sont les travaux déjà réalisés dans la littérature.

5.1. Etude du bruit à basse fréquence dans la littérature

Le bruit RTN est un phénomène de plus en plus étudié pour les mémoires OxRAM. En effet, il y a deux intérêts principaux à l'étudier. Le premier est simplement qu'il peut induire des problèmes de fiabilité, et notamment de lecture d'état résistif. Il convient donc de l'étudier pour mieux le comprendre et pouvoir ainsi le réduire. Le second intérêt est que le RTN est un moyen assez puissant pour sonder ce qu'il se passe à l'échelle microscopique et de détecter l'impact de pièges uniques sur un signal électrique.

Le bruit RTN consiste en des oscillations du courant, le plus souvent entre deux valeurs, à des basses fréquences (de l'ordre de grandeur du Hz), comme représenté en Figure IV.28, pour des mémoires Flash NOR [15].

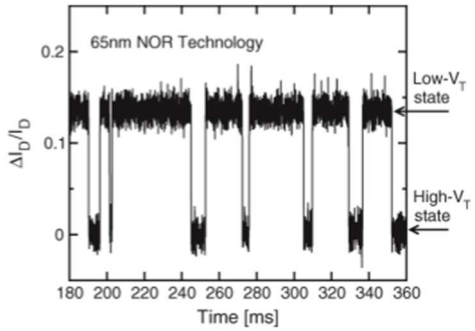


Figure IV.28 : Observation de bruit à basse fréquence sur des mémoires Flash NOR [15].

Ce phénomène, très largement étudié pour les technologies MOS [15, 16, 17, 18, 19], est rattaché à des évènements de capture et d'émissions d'électrons par des pièges présents dans les dispositifs.

Plus récemment, ce phénomène a été mis en lumière pour les mémoires résistives, et accusé d'entrainer des échecs lors des opérations de lecture [20, 21, 22, 23]. En Figure IV.29, un exemple de lecture de bruit RTN sur des OxRAM à base de hafnium [24].

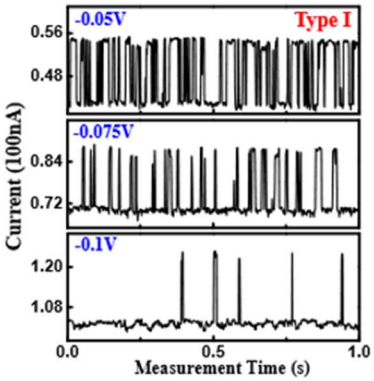


Figure IV.29 : Trois mesures de bruit à basse fréquence sur des dispositifs OxRAM [24].

Comme pour les dispositifs MOS, le bruit RTN chez les OxRAM est dû à des pièges et des phénomènes de captures et d'émissions d'électrons [20, 24, 25, 26, 27, 28]. Cependant, la nature exacte de ces pièges prête encore à débat. Dans [29], F.M Puglisi et al. suggèrent deux principales hypothèses pour expliquer ce bruit à basse fréquence dans les mémoires résistives :

- La première est liée au phénomène de blocage de Coulomb qui va interférer avec la conduction TAT. Comme nous l'avons déjà expliqué, la conduction est supposée être due à des lacunes d'oxygène présentes, par lesquelles circulent les électrons de conduction. Or il est possible que des oxygènes interstitiels soient situés proche de ces lacunes d'oxygène (lors de la génération de lacunes d'oxygène, un oxygène interstitiel est également généré). Ces atomes d'oxygène sont ensuite susceptibles de piéger des électrons. Un défaut ainsi chargé électriquement va générer un champ électrique qui pourra interagir avec le courant, et le bloquer. Les fluctuations de courant seraient ici liées aux phénomènes de charge et de décharge des oxygènes interstitiels : une oscillation vers le bas indique une capture d'électron, tandis qu'une oscillation de courant vers le haut indique l'émission d'un électron par cet oxygène. La Figure IV.30 illustre ce phénomène de blocage de Coulomb pour les mémoires résistives, à base d'oxyde.

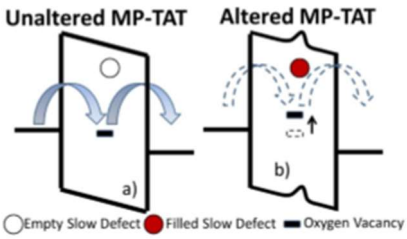


Figure IV.30 : Illustration du blocage de Coulomb. (a) Conduction TAT "normale", en l'absence de charges à proximité. (b) Perturbation de la conduction TAT par la capture d'un électron par un oxygène interstitiel [29].

- La seconde est liée à des différents états des lacunes d'oxygènes, stables et métastables. Ce phénomène n'a pas encore été démontré dans l'oxyde d'hafnium, mais sur de l'oxyde de silicium [30]. Selon cette hypothèse, une lacune d'oxygène chargée positivement pourrait avoir deux états possibles : un état métastable qui résulte en une rapide transition vers un état neutre (et inversement), dû à une faible barrière énergétique, et un état stable, séparé de l'état métastable par une barrière énergétique d'environ 0.8eV. Le courant TAT serait lié aux captures et émission rapide d'électron par les états métastables, tandis que le RTN serait lié au piégeage et dépiégeage par les états stables, moins rapides. Ce principe de trois états de lacunes d'oxygène est décrit en Figure IV.31.

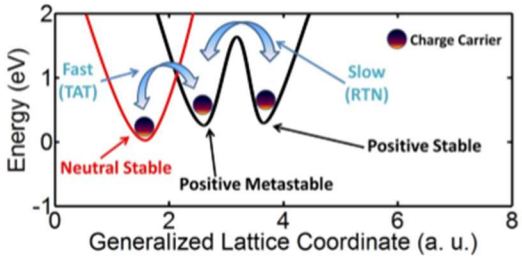


Figure IV.31 : Configuration avec trois états de lacunes d'oxygène. La transition d'un état métastable vers l'état neutre est liée au phénomène de courant TAT. Les fluctuations de courant de type RTN seraient quant à elles liées aux transitions entre l'état métastable et l'état positif stable. On observerait une chute de courant lors du passage vers l'état stable et une ré-augmentation de celui-ci lors de la transition vers l'état métastable [29].

Les fluctuations de courants causées par le bruit RTN peuvent représenter un véritable problème de fiabilité, puisque l'état résistif de la mémoire change dans le temps (en fonction des phénomènes d'émission et de capture). C'est pourquoi nous avons décidé, pour la dernière partie de ce travail de thèse, d'étudier le bruit RTN, afin de trouver des pistes pour le réduire, notamment en explorant l'impact du courant de forming. En effet nous avons vu qu'un courant de forming élevé permettait d'obtenir un filament plus épais, même après cyclage. Il semble donc intéressant d'évaluer l'effet de cet épaississement du filament conducteur sur le bruit à faible fréquence.

5.2. Set-up expérimental de l'étude du bruit RTN

Evaluer le bruit RTN est une manipulation sensiblement différente à celle effectuée jusque-là : il convient, pour cela, d'analyser des mesures de courant, sur des temps de plusieurs secondes, ce qui rend ce travail impossible sur des matrices entières, pour des questions évidentes de durée et de quantité de données à traiter. C'est pourquoi, nous avons ici utilisé le

véhicule de test MARS, utilisé dans le second chapitre de ce rapport de thèse, avec le banc de test associé.

Une mesure de RTN consiste en une lecture du courant, sur une durée de plusieurs secondes. Le choix de la durée de test et du temps d'échantillonnage est important, puisqu'elle définit la fréquence du bruit RTN que l'on veut détecter. Ici nous nous intéressons au bruit basse fréquence. Nous avons choisi de tester les cellules sur une durée de 180 secondes, avec un temps d'échantillonnage de 45ms, ce qui nous permet d'accéder à des fréquences maximales de 22Hz. Ainsi, après forming et quelques cycles (réalisés ici en quasi-statique pour mieux cibler les états de résistance que nous souhaitions), une lecture à 0.1V est réalisée pendant 180 secondes, par un SMU. Pour chaque cellule, on obtient alors une courbe de courant en fonction du temps, que l'on peut alors analyser.

5.3. Analyse du bruit RTN

Une fois qu'une courbe RTN est obtenue, il faut la traiter et à en retirer les informations qui nous intéressent. L'une des informations les plus pertinentes est le nombre de niveaux de courant que l'on peut observer. S'il n'y a qu'un seul niveau de courant, il n'y a pas de bruit RTN. En revanche, s'il y a plusieurs niveaux, il convient de les compter, pour se faire une idée du nombre de défauts. En Figure IV.32, nous avons représenté une mesure de bruit brute, présentant un grand nombre de niveaux de courants.

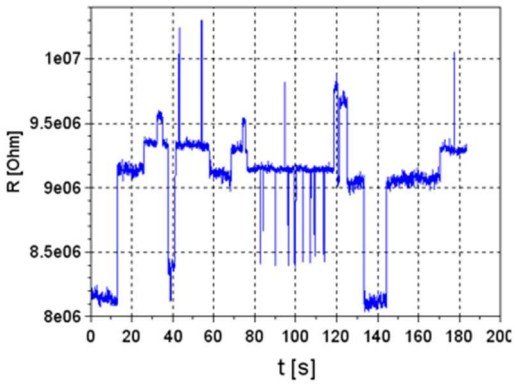


Figure IV.32 : Mesure de bruit RTN, à 0.1V mesurée sur un dispositif T3 du véhicule MARS, après un forming à 400µA. Ici nous avons traduit le courant mesuré en résistance.

Plusieurs techniques permettent de compter le nombre de niveaux. La première consiste simplement à compter le nombre de pallier de courant, directement sur la courbe du courant (ou de la résistance). Parfois cette technique est suffisante, mais ce n'est, par exemple, pas le cas ici, où il est difficile de discerner certains niveaux. Une autre technique, très utilisée est de tracer un histogramme de la distribution des niveaux de résistance, tel que représenté en Figure IV.33 pour la même mesure.

Cette technique a le mérite de représenter très clairement les différents niveaux. Il faut néanmoins s'appliquer à choisir un pas en résistance/courant convenable, pour ne pas "rater" certains niveaux.

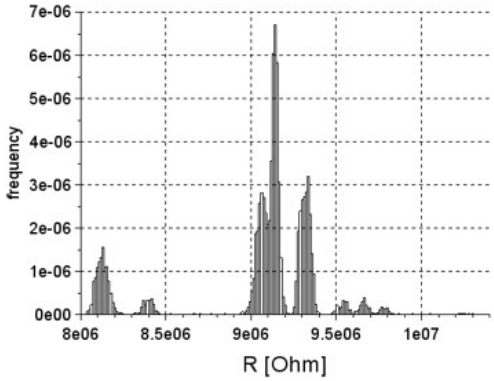


Figure IV.33 : Histogramme des niveaux de résistances, pour la mesure de courant de la Figure IV.32.

Le problème de cette technique est qu'elle ne permet pas de voir clairement les niveaux très rapides. Par exemple, sur la figure IV.32, on distingue un niveau aux alentours de 10.2MΩ, très furtif. En effet, il semble être caractéristique d'un piège doté d'un taux d'émission très court. Un électron capturé par ce piège est très vite réémis, ce qui rend ce piège difficile à détecter. C'est pourquoi, on ne le distingue quasiment pas sur l'histogramme.

A l'opposé, la troisième technique est la plus sensible aux niveaux très furtifs. Cette technique est présentée dans [26]. Celle-ci consiste à tracer, pour chaque n^{ieme} point mesuré, le point (n+1) suivant. Une telle représentation est illustrée en Figure IV.34, toujours pour la même mesure de bruit.

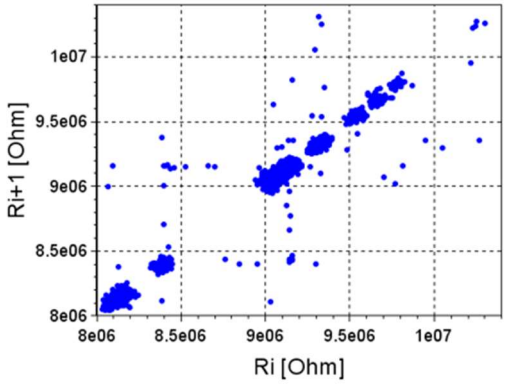


Figure IV.34 : tracé de chaque point mesuré en fonction du point précédent

Sur une telle courbe, il faut lire les points présents sur les diagonales : ils représentent les mesures pour lesquels le $n^{ième}$ et le $(n+1)^{ième}$ point sont identiques. Ainsi, n'importe quel niveau de courant existant pour au moins deux mesures consécutives sera visible sur cette courbe. Par exemple, le niveau à 10,2MΩ dont nous parlions ci-dessus est facilement visible

ici. Les points ailleurs que sur la diagonale correspondent aux transitions entre les différents niveaux. Cela donne les formes rectangulaires qui indiquent deux niveaux entre lesquels il y a beaucoup de transitions.

La faiblesse de cette technique est qu'elle différencie assez mal deux niveaux proches. Il est donc très souvent nécessaire de croiser ces trois techniques de mesures pour avoir une meilleure idée du nombre de niveaux de bruit. C'est pourquoi le traitement des courbes de bruits est un travail assez long. Ici, on peut considérer que l'on distingue 9 niveaux de bruits différents. Pour rappel, nous avons ici pris en exemple une mesure particulièrement riche en niveaux de courants.

Enfin, nous analyserons l'amplitude du bruit normalisée (c'est-à-dire en $\Delta R/R$).

5.4. Impact du courant de forming sur le nombre de niveaux de courant

Afin de comprendre l'impact du courant de forming sur le bruit RTN, nous avons décidé de commencer par mesurer son impact sur le nombre de niveaux de courant. Pour cela, nous avons formé 4 groupes de 35 cellules, à 4 courants de forming différents : 50µA, 125µA, 200µA et 350µA. Tous les formings ont été réalisés avec une tension de 3.5V sur l'électrode supérieure. Les cellules ont ensuite subi une opération de reset en quasi-statique, en s'assurant que les quatre échantillons présentaient les mêmes ordres de grandeurs de résistance. En effet, nous voulons voir l'impact du forming, et le distinguer de celui de l'état résistif. Cependant, lors du traitement des données, nous avons volontairement séparé les résultats en fonction de la résistance des cellules pour voir l'impact de la résistance. En effet certains papiers, dans la littérature, démontrent des fluctuations de courants plus importantes à mesure que la résistance augmente [25, 31, 32].

La courbe représentant le nombre de niveaux de courant moyen en fonction du courant de forming est présentée en Figure IV.35. Nous avons tracé trois courbes, pour séparer les résistances inférieures à 200kΩ, inférieures à 500kΩ et supérieures à 500kΩ. Ces trois différents états résistifs ont été obtenus en modulant la tension de reset (1.2V, 1.4V et 1.6V, en quasi-statique).

On constate que, lorsque le courant de forming utilisé augmente, on observe une augmentation du nombre moyen de niveaux de résistances détectés. Ceci est valable quel que soit le niveau de résistance choisi : les trois courbes présentent le même impact du courant de forming. En effet, les mesures précédentes et les comparaisons avec notre modèle nous ont

appris que le courant de forming contrôle le rayon du filament. De plus, si, à résistance équivalente, le rayon du filament est supérieur, il est également possible que la zone de gap soit plus grande (afin de conserver une résistance du même ordre de grandeur). On a donc, lorsque le courant de forming augmente, une zone de gap d'un volume supérieur. Le nombre de défauts (oxygènes interstitiels ou lacunes d'oxygène) dans cette zone est donc logiquement plus élevé. Il est donc logique d'observer plus de niveaux de courant. Un schéma de l'évolution de la géométrie du filament en fonction du courant de forming et ses impacts sur le nombre de défauts, ainsi que des courbes de bruits correspondantes est représenté en Figure IV.36.

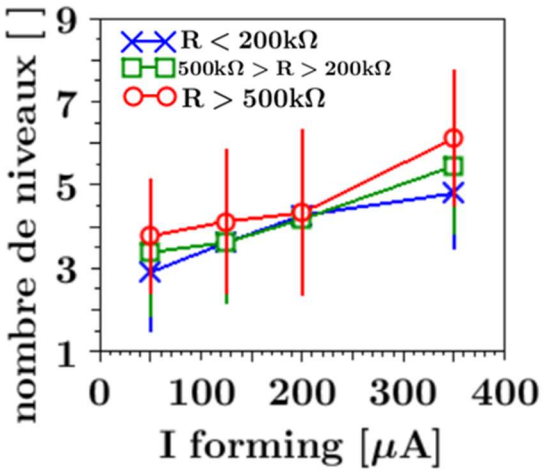


Figure IV.35 : Evolution du nombre de niveaux en fonction du courant de forming, pour différents états de résistances.

On observe, dans un second temps que lorsque la résistance augmente, le nombre moyen de niveau de résistances augmente légèrement. Cet impact est moins important que celui lié au courant de forming. En effet, l'augmentation de la résistance peut être lié à l'augmentation de la zone de gap. Ceci augmente donc possiblement le nombre de défauts. Cependant, ici l'augmentation de la zone de gap ne se fait que dans une dimension, alors que dans le cas du rayon du filament, celle-ci se fait dans deux dimensions. C'est pourquoi l'impact est ici moins important.

Ces mesures réalisées sur le nombre de niveaux de résistances observé sont cohérentes avec une augmentation du rayon du filament conducteur lorsque le courant de forming augmente. En revanche, d'un point de vue de la fiabilité des dispositifs, le nombre de niveaux n'est pas l'information la plus pertinente à relever. En effet, deux niveaux très éloignés sont par exemple un problème plus important que trois niveaux très proches, qui n'induiront pas d'erreurs de lecture très graves. C'est pourquoi, nous avons également décidé de regarder l'impact qu'a le courant de forming sur l'amplitude normalisée des oscillations de courant.

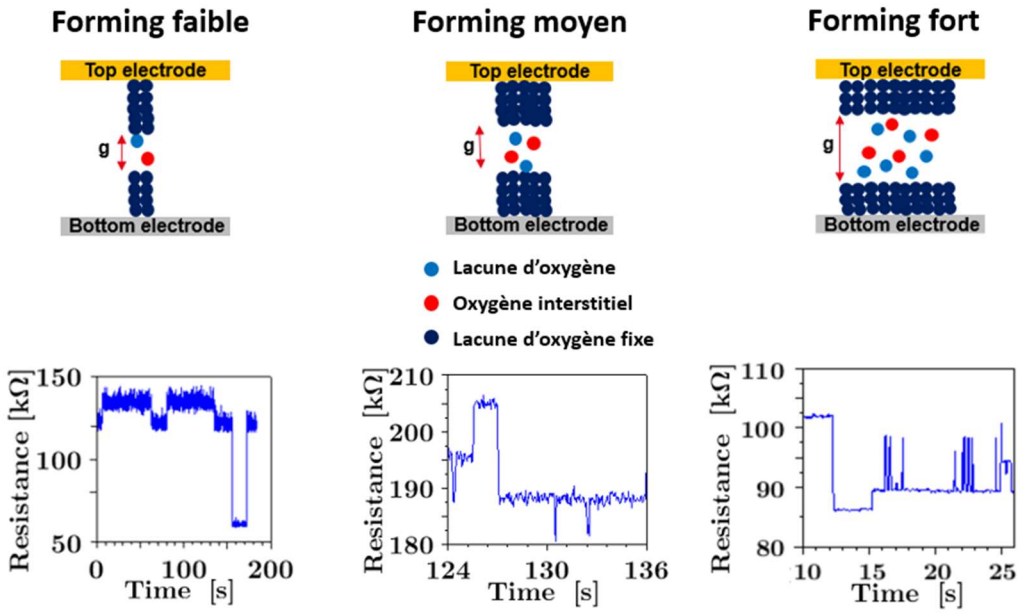
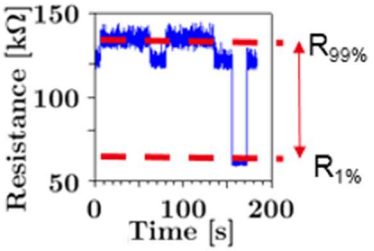


Figure IV.36 : Illustration, pour trois différentes intensités de courants de forming (50, 200 et 250µA), de l'état du filament (avec les deux types de défauts qu'on peut y trouver et qui peuvent générer des fluctuations de courant) et trois exemples de mesures de bruit correspondantes. Un fort courant de forming induit un filament conducteur plus large, dans le lequel il est susceptible d'y exister plus de défauts, aussi bien des oxygènes interstitiels que des lacunes d'oxygène.

5.5. Impact du courant de forming sur l'amplitude du bruit

Pour cela, nous avons repris les 4 jeux de 35 courbes obtenues. Il nous faut maintenant définir ce que l'on entend par amplitude des oscillations de courant. Nous avons choisi de définir l'amplitude par l'écart, en échelle logarithmique, entre les 1% de résistance les plus faibles et les 1% de résistance les plus élevées, mesurés sur une même courbe (cf. Fig. IV.37). Pour rappel, chaque mesure contient 4000 points (une mesure toutes les 45ms, pendant 180s). 1% des points représente donc 40 points de mesures.



$$Amplitude = log\left(\frac{R_{99\%}}{R_{1\%}}\right)$$

Figure IV.37 : Définition de l'amplitude des fluctuations de résistance

En plus de reprendre les courbes précédemment mesurées, nous avons ajouté :

- Une série de 50 mesures à un courant de forming de 75µA. En effet, l'allure de la courbe obtenue nécessitait un point supplémentaire, à ce niveau-là.

- Une série de 120 mesures, pour un état LRS (défini ici comme inférieur à 10kΩ), soit 30 mesures par valeur de courant de forming. Ces états LRS ont été obtenus avec des opérations de set en quasi-statique (pour s'assurer un meilleur contrôle des états obtenus), à un courant de 50µA.
- Une série de 80 mesures, pour un état intermédiaire entre HRS et LRS (dont la résistance est comprise entre 10kΩ et 30kΩ), soit 20 mesures par valeur de courant de forming. La raison pour laquelle moins de points ont été pris ici est que cet état était relativement difficile à cibler.

Les deux dernières séries ont pour but de comparer nos résultats avec certains papiers de la littérature détaillant des comportements différents entre les différents états résistifs [25, 31, 32, 33]. Dans ces papiers, il apparaît que le régime LRS est bien moins influencé par les fluctuations RTN que le régime HRS.

La courbe détaillant l'évolution de l'amplitude moyenne des fluctuations de résistance est représentée, pour les trois états (HRS, intermédiaire et LRS), en Figure IV. 38.

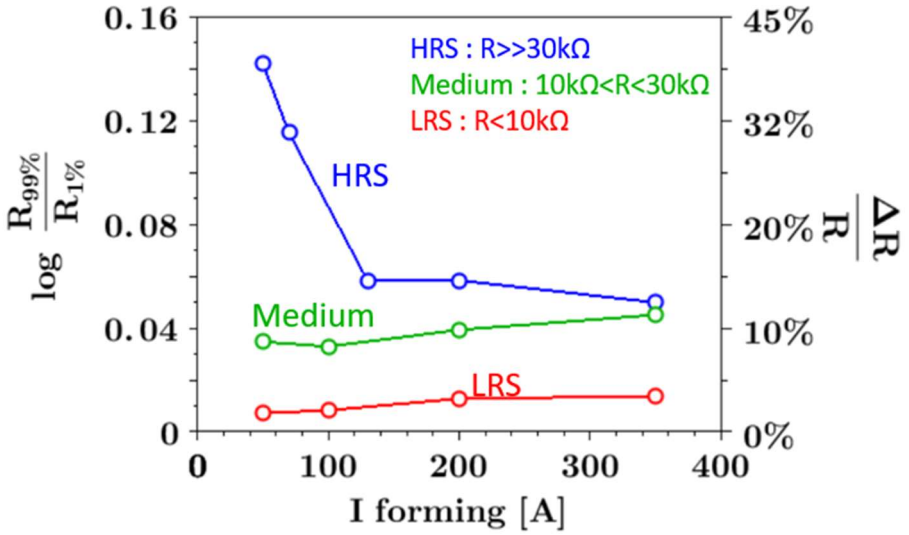


Figure IV.38 : Impact du courant de forming sur l'amplitude moyenne des oscillations de résistance, pour différents états résistifs.

On constate tout d'abord que les trois niveaux résistifs sont très clairement distincts. Alors que l'état HRS est clairement affecté par le courant de forming (nous y reviendrons plus tard), les états intermédiaires et faiblement résistifs ne sont visiblement pas affectés. De plus, on voit bien qu'à mesure que la résistance augmente (en passant d'un état LRS à intermédiaire, ou d'intermédiaire à HRS), l'amplitude des fluctuations augmente sensiblement. Ceci confirme les résultats de la littérature. Le régime HRS se comporte différemment vis-à-vis des

perturbations RTN. Lorsque le filament conducteur est intact, le courant est déjà bien établi dans la mémoire, et l'impact d'un piège n'affecte que peu le niveau de courant.

Il convient maintenant de nous concentrer sur la courbe bleue, à propos de l'état hautement résistif. Sur cette courbe, il est clair que l'augmentation du courant de forming abaisse nettement l'amplitude du bruit. Cette diminution du bruit, se fait principalement entre 50 et 100µA (d'où l'ajout des 50 mesures à 75µA, pour rajouter un point au niveau du point d'inflexion de la courbe). Il est important d'insister sur les valeurs en jeu. En effet, aux alentours de 50µA, l'amplitude du bruit ici en échelle logarithmique à environ 0.15, représente un écart de 40% entre les 1% de résistances les plus bas et les 1% les plus hauts. Il est clair qu'une telle différence impacte fortement la fiabilité des dispositifs. De plus, il s'agit d'une valeur moyenne. Comme en atteste la figure 37, certaines mesures présentent une amplitude de bruit encore plus impressionnante. Il est donc très intéressant de remarquer qu'une augmentation du courant de forming vers 100 ou 150µA permet d'abaisser nettement cette amplitude. Cette intensité permet également de trouver un compromis au sujet du nombre de niveau obtenu : il n'est pas nécessaire de monter le courant le forming au-delà de 200µA, puisque l'amplitude ne diminue plus au-delà. C'est, de plus, intéressant, d'un point de vue énergétique, même si l'étape de forming n'est pas l'étape sur laquelle la dépense énergétique est décisive (le forming n'est réalisé qu'une seule fois dans la vie de la cellule). En Figure IV.39, nous avons repris la courbe de la Figure IV.38, pour les états hautement résistifs, et avons illustré la courbe de deux mesure de courant, après des formings de 50µA et de 350µA, représentatives des valeurs indiquées par la courbe, afin de mieux se rendre compte de la différence d'amplitude mise en jeu ici.

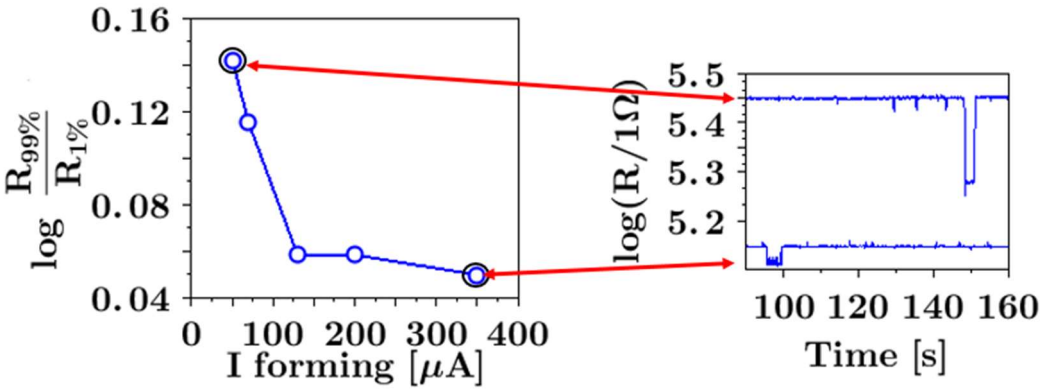


Figure IV.39 : Illustration de la différence d'amplitude après un forming fort, et après un forming faible. On peut noter que cette différence est significative.

Là encore, il convient de relier cet impact qu'a le courant de forming sur l'amplitude des fluctuations de courant avec le lien entre le courant de forming et le rayon du filament, que nous avons analysé précédemment dans ce chapitre. Il semble ainsi, qu'un filament épais, obtenu

après un forming d'une forte intensité, conduise à un abaissement des perturbations de courant RTN, tandis qu'un filament plus fin y est plus sensible. On peut expliquer cela physiquement. Comme nous l'avons vu, les perturbations RTN sont dues à des défauts présents dans la zone de rupture du filament. Ces défauts, qui peuvent être des lacunes d'oxygènes dans un état différent de celui permettant le courant TAT ou des oxygènes interstitiels capables de capturer des électrons. Dans les deux cas, ces perturbations sont causées par la capture d'électrons par ces défauts. En Figure IV.40, nous avons schématisé deux filaments conducteurs (l'un fin et l'autre épais) et imaginé une distribution de défauts dans le gap. On peut noter que nous avons pris un gap plus grand pour le cas d'un filament épais. En effet, si le filament est épais, la résistance a tendance à diminuer. Donc pour obtenir une résistance comparable à celle du filament fin, nous avons élargi la zone de rupture du filament. En bleu clair sont représentées les lacunes d'oxygène responsable du TAT et en rouge les défauts responsable du RTN. Nous avons alors supposé qu'un défaut est chargé (celui-ci cerclé en noir). Ce défaut chargé va alors émettre un champ électrique, responsable de la perturbation du courant. Il est alors assez intuitif de voir que dans le cas d'un filament fin, la présence d'une perturbation va interagir avec une plus large portion du filament conducteur que dans le cas d'un filament épais. En effet, dans le cas de ce dernier, même si un défaut est chargé, il reste un plus grand nombre de chemins possibles pour le courant. Pour un filament fin, le courant passe nécessairement à proximité du défaut chargé électriquement. C'est pourquoi, le courant est plus fortement perturbé. Au contraire un filament large encaisse plus facilement les perturbations du champ électrique.

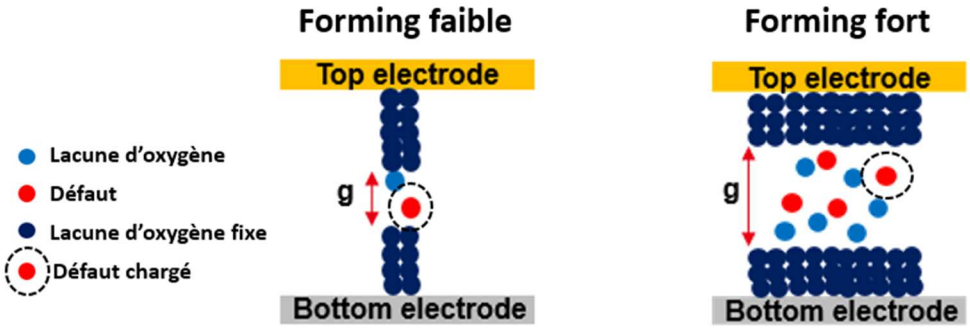


Figure IV.40 : Différence microscopique après un faible et un fort forming. Impact de la présence d'un défaut chargé sur l'état de conduction.

Cette dernière manipulation confirme un peu plus la thèse selon laquelle le forming a un impact sur le rayon du filament conducteur. De plus, nous commençons à apercevoir un intérêt à utiliser un fort forming. Puisque cela réduit l'amplitude du bruit, on peut supposer qu'il va y avoir un impact sur la fiabilité des dispositifs. En effet, si au cours d'une même mesure de bruit, la résistance d'une cellule évolue de façon importante, comme c'est le cas avec les mesures

après un faible courant de forming, on peut supposer que d'une lecture à une autre, on est susceptible de lire un état différent. C'est ce que nous allons étudier dans la partie suivante.

5.6. Impact du courant de forming sur des lectures successives

Pour attester de l'impact du forming, et de l'amplitude du bruit RTN qui en résulte, sur la fiabilité des dispositifs, nous allons recentrer l'étude sur une vision plus statistique. Nous avons ainsi choisi de préparer 3 groupes de cellules différents, selon trois courants de forming : 50, 130 et 350µA. Après forming, ces cellules ont subi un effacement pour les mettre en état HRS. Ensuite, ces cellules ont subi chacune 1800 lectures successives, c'est-à-dire sans opération d'écriture ou d'effacement. Ainsi, dans ce contexte, nous attendons d'une mémoire "parfaite" que les 1800 lectures soient identiques (or incertitudes de lecture liés au montage).

On peut ensuite tracer pour chaque cellule la distribution des 1800 lectures. Nous voulions comparer la dispersion de ces lectures pour les trois groupes. Comme les distributions ne sont pas toutes gaussiennes, comme nous le verrons, nous avons donc choisi de ne pas simplement mesurer l'écart type, mais l'écart entre -2σ et +2σ (mesuré sur les distributions à échelle de déviation standard), qui représente 95% de la masse totale des mesure. En Figure IV.41 (a), nous avons représenté, pour chaque cellule, cet écart entre -2σ et +2σ, en fonction de la résistance moyenne mesurée. La valeur moyenne de cet écart, pour chaque valeur de courant de forming est tracée en Figure IV.41 (b). A noter que 3 temps de lectures différents ont été testés (10µs, 100µs et 1ms). Mais comme nous n'avons pas constaté de différence entre les trois, nous n'avons tracé ici que les 1800 lectures à 10µs.

On remarque tout d'abord qu'il n'y a pas d'impact clair de la résistance sur la dispersion. Ensuite, il est net que les mesures réalisées après le forming à bas courant sont plus dispersées. Sur la Figure IV.41 (a) les points en bleus sont très visiblement plus hauts que les points verts et rouges. Cela ressort très facilement sur la Figure IV.41 (b), où la valeur moyenne de l'écart à 2σ (en échelle logarithmique, toujours) est bien plus élevée pour 50µA que pour les autres. On retrouve par ailleurs la forme de la courbe représentée en Figure IV.38, pour les états HRS, avec notamment la grande variation entre 50 et 130µA. Ceci n'est pas une coïncidence : cela montre que des mesures de lectures successives sont équivalentes à des mesures de bruits continues. Cela démontre l'impact sur la lecture d'une seule cellule du bruit RTN : d'une lecture à une autre, on pourra lire deux valeurs de résistances différentes, et de façon plus affirmée à mesure que le courant de forming utilisé était faible.

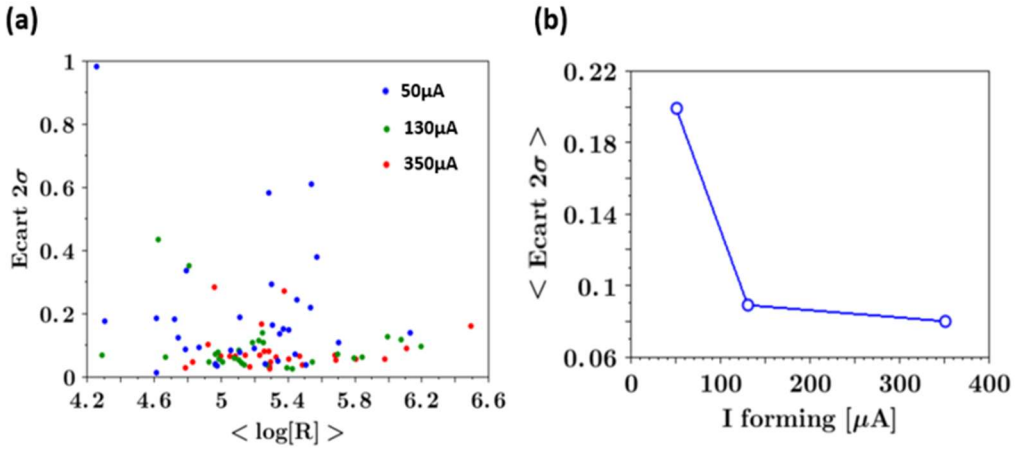


Figure IV.41 : (a) Ecart à 2σ en fonction de la résistance, pour trois courants de forming différents. (b) Tracé de la valeur moyenne de cet écart pour les trois courants de forming.

En Figure IV.42 (a), nous avons montré un exemple d'une mesure de distribution sur une cellule formée à 50µA. Cette mesure présente des oscillations de mesures de résistances particulièrement impressionnantes et démontre aisément à quel point le bruit RTN peut interférer dans les mesures de résistances. De plus pour confirmer que l'origine de ces perturbations de mesures de résistances est due au bruit RTN, nous avons tracé en Figure IV.42 (b), les 1800 mesures de résistances par ordre chronologique. On constate que la courbe obtenue est très similaire à une courbe de mesure de bruit. Les changements de résistances ne sont par exemple pas une dégradation avec le temps de la résistance, mais bien des évènements de piégeages et dépiégeages.

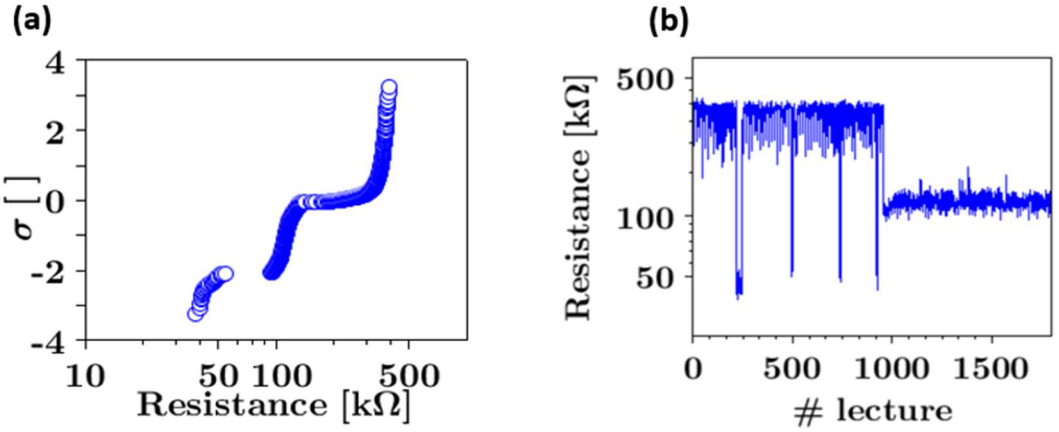


Figure IV.42 : (a) mesure d'une distribution de lectures successives sur une cellule formée à 50µA. (b) tracé de cette distribution par ordre chronologique des mesures de lectures.

Les dernières mesures ont montré, sans équivoque, que sur une seule cellule, d'une lecture à une autre, le bruit RTN pouvait conduire à d'importantes erreurs de lectures. Sur la cellule prise en exemple sur la Figure IV.42, la même cellule peut être lue comme ayant une résistance de 50kΩ, ou presque 400kΩ, ce qui chez certaines mémoire peut suffire à distinguer

un état LRS d'un état HRS. Le bruit RTN a donc un très net impact sur les dispersions temporelles. Il est donc intéressant maintenant de se demander si ce bruit a également un impact sur les dispersions spatiales. Intuitivement, on peut penser que oui : mesurer la dispersion cellules à cellules, de cellules qui elles même sont dispersées d'une mesure à une autre, peut probablement conduire à une plus grande dispersion.

5.7. Impact du courant de forming sur des lectures de matrices 4kbits

Pour ce dernier test sur le bruit RTN chez les OxRAM, nous avons repris le véhicule de test MAD, et le banc de test Electroglas associé. Nous avons choisi de mesurer des matrices 4kbits, sur des dispositifs classiques TiN/Ti10nm/$HfO_2$10nm/TiN. Les transistors présents sur MAD sont différents de ceux sur MARS et nous ne pouvons pas monter ici sur des courants de 200 et 300µA. Nous avons donc choisi trois groupes de dispositifs : 50, 80 et 150µA. Cependant, ceci ne semble pas être un problème, puisque nous avons vu que les changements de comportement et d'amplitude des variations de courant se font entre 50 et environ 150µA : au-delà, l'amplitude ne varie plus beaucoup (cf. Fig. IV.38 et IV.41). Après le forming, ces trois matrices ont subi deux cycles d'écriture/effacement, en pulsé. Toutes les opérations de lectures ont été réalisées avec un courant de 50µA.

Pour le premier test, les trois matrices ont subi exactement les mêmes opérations de reset, afin que la seule différence soit le forming. Nous avons également lu les matrices à 3 temps de lecture différents : 10µs, 100µs et 1ms. Nous n'avons cependant pas observé de variations avec le temps de lecture (cf. Fig. IV.43).

Comme nous n'avons pas observé de différences, nous n'avons traité dans la suite que les lectures de 10µs. La comparaison des distributions des trois matrices est représentée en Figure IV.44.

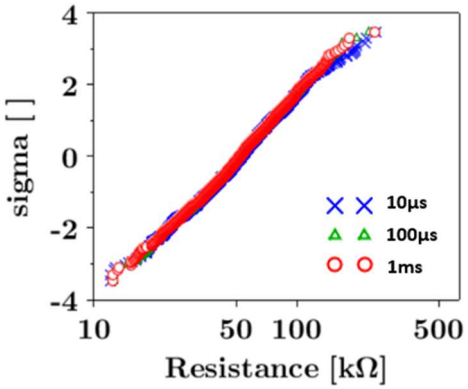


Figure IV.43 : Comparaison des différents temps de lecture, pour la matrice 4kbit ayant subi un forming de 80µA.

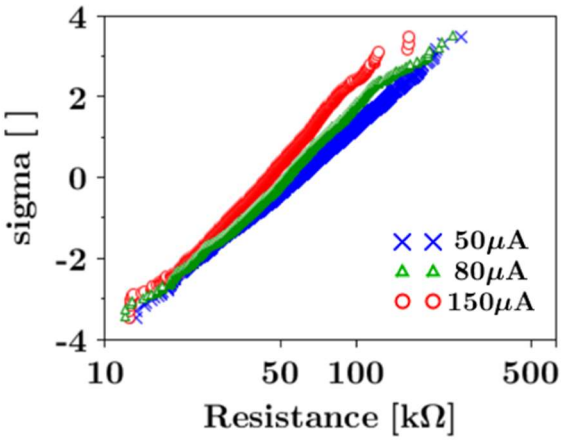


Figure IV.44 : Comparaison des distributions sur les trois matrices, formées à 3 courants de forming différents.

On constate un impact assez net du courant de forming. Logiquement les valeurs moyennes obtenues sont moins bonnes après un fort courant de forming, puisque le reset a plus de mal à effacer un filament épais. En revanche, on remarque bien que la dispersion des résistances, après un fort forming, est moins grande. Malgré une résistance moyenne moins satisfaisante, la courbe à 150µA est, aux alentours de -3σ, au même niveau que celle à 50µA. Il semble donc que l'utilisation d'un forming d'intensité faible résulte bien en une dispersion plus importante des distributions de résistance.

Nous voulions désormais comparer des distributions à valeurs moyennes équivalentes, pour les trois valeurs de courant de forming. Pour ce faire, nous avons pris trois nouvelles matrices et leur avons fait subir le même processus, à une différence près : les courbes à 50, 130, et 150µA ont subi des opérations de reset incrémentée de 0.1V, afin de décaler les distributions, pour cibler des résistances moyennes équivalentes. Il est logique, en effet, de réaliser des effacements plus intenses, lorsque le filament est plus épais.

Les trois distributions obtenues à l'issu de ces mesures sont représentées en Figure IV.45.

On constate tout d'abord que le fait d'augmenter légèrement le reset pour les matrices à fort courant de forming permet bien d'aligner les trois distributions sur les mêmes valeurs moyennes. On remarque ainsi très facilement les différences en matière de dispersion : la distribution obtenue après un forming à fort courant est nettement moins dispersée que celle obtenue après un forming à faible courant (comparaison entre les courbes rouge et bleue). Il est difficile de dire quelle est la part due à l'augmentation de la tension de reset. Cependant, le fait que la courbe rouge soit moins dispersée aussi bien du côté des faibles que des fortes résistances laisse penser que ce resserrement de la dispersion n'est pas due à l'augmentation de la tension de reset. En effet, il n'y pas de raison que l'augmentation de reset resserre les distributions

(surtout du côté des fortes résistances). Ceci est confirmé par l'étude réalisée dans le second chapitre de ce rapport de thèse (en Figure II.46) : l'augmentation de la tension de reset augmente même légèrement l'écart type des distributions. Nous avions alors déjà remarqué que cette augmentation de la tension de reset impactait très majoritairement la valeur moyenne des distributions. On peut donc légitimement supposé que cette amélioration de la dispersion pour les matrices formées à fort courant, n'est pas due à l'augmentation de la tension de reset, mais bien à l'épaississement du filament, liée à l'opération de forming.

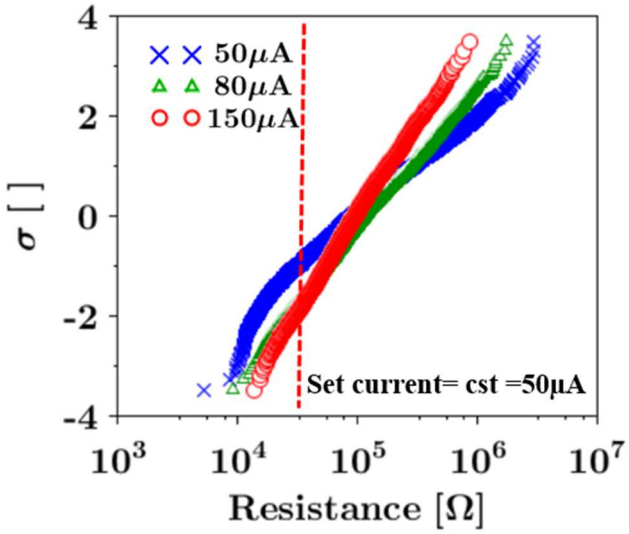


Figure IV.45 : Evolution des distributions HRS en fonction du courant de forming employé, après ajustement des valeurs moyennes.

Grâce à ces courbes de distributions à échelle de déviation standard, il est facile de quantifier les différences entre la matrice formée à fort courant et celle formée à faible courant. Sur la Figure IV.45, nous avons indiqué une ligne démarquant les résistances inférieures à 20kΩ. Ainsi, après un forming de 150µA, seules 2-3% (-2σ) des cellules ont une résistance inférieure à 20kΩ, tandis que cela concerne plus de 15% (-1σ) des cellules formées avec un courant de 50µA. Augmenter le courant de forming semble donc être bénéfique pour améliorer la fenêtre résistive.

Cette amélioration de la dispersion est également positive si on vise des applications "multi-level", où la précision de la valeur de la résistance est une donnée essentielle.

6. Conclusion

Ce dernier chapitre de ce manuscrit de thèse avait pour but de réaliser la transition entre une étude sur le fonctionnement d'un dispositif unique à celui d'une matrice de plusieurs kbits. Nous avons donc présenté le nouveau véhicule de test MAD adapté aux matrices OxRAM, ainsi que le banc de test associé. Comme nous l'avions déjà annoncé dans les parties précédentes, nous avons remarqué que ces tests sur des matrices mémoires sont absolument nécessaires dans le cas des OxRAM, pour étudier leur variabilité, qui reste leur principal point faible.

Nous avons ensuite confronté notre modèle à ces mesures de statistiques. Le choix d'utiliser un modèle simple semi-analytique nous permet ainsi de lancer un grand nombre de simulations, sur des durées qui restent raisonnables. Nous avons donc pu fitter les distributions de tensions de set en fonction de la vitesse de rampe, en introduisant une distribution gaussienne de la valeur de longueur de gap.

Dans un troisième temps, nous avons analysé l'impact du courant de forming, après cyclage, sur les valeurs de tensions de set. Nous avons ainsi observé que l'augmentation du courant de forming résultait en des tensions de set plus faibles. En nous basant sur les travaux présents dans la littérature, et sur notre modèle physique, nous avons relié le courant de forming au rayon du filament conducteur. Cette hypothèse a ensuite été confirmée par les mesures de bruits qui démontrent à la fois un plus grand nombre de niveaux de courant et une amplitude des fluctuations plus faibles, lorsque le courant de forming augmente, ce qui est cohérent avec une augmentation de l'épaisseur du filament conducteur.

Enfin, nous avons vu que cette baisse d'amplitude des fluctuations du bruit RTN résultait en des distributions moins dispersées, à la fois en temporel (d'une lecture à une autre, sur une même cellule) et en spatial (d'une cellule à une autre, sur des matrices 4kbits). Ainsi, d'un point de vue de la variabilité, augmenter le courant de forming semble être bénéfique.

7. Références du chapitre IV

[45] A. Fantini et al., "Intrinsic switching variability in HfO_2 RRAM", 5th International Memory Workshop (IMW), 2013

[46] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy and D. Ielmini, "Statistical Fluctuations in HfOx resistive-switching memory: part I – Set/Reset variability", IEEE Transactions on Electron Devices, vol. 61, no. 8, august 2014

[47] R. Degraeve et al., “Causes and consequences of the stochastic aspect of filamentary RRAM”, Microelectronic Engineering 147, 171–175, 2015

[48] D. Garbin et al., “Modeling of OxRAM variability from low to high resistance state using a stochastic trap assisted tunneling-based resistor network”, Eurosoi ULIS pp.125-128, 26-28 Jan. 2015

[49] L. Kadura, "Electrical characterization of Resistive-RAM Kbits devices: Instrumental development and statistical analysis", Internship report, September 2015

[50] http://literature.cdn.keysight.com/litweb/pdf/5989-2785EN.pdf?id=629794

[51] http://www.atmel.com/Images/Atmel-2549-8-bit-AVR-Microcontroller-ATmega640-1280-1281-2560-2561_datasheet.pdf

[52] http://www.electroglas.com/PDF/EG4090u+.pdf

[53] C. Nguyen, C. Cagli, L. Kadura, J-F. Nodin, S. Bernasconi and G. Reimbold, "A New Test Vehicle For RRAM Array Characterization", 30th International Conference on Microelectronic Test Structures (ICMTS), 2017.

[54] C. Cagli, J. Buckley, V. Jousseaume et al., "Experimental and theoretical study of electrode effects in HfO2 based RRAM," IEEE Electron Devices Meeting IEDM, pp. 658–661, 2011

[55] D. Ielmini, F. Nardi and C. Cagli, “Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories”, Nanotechnology 22, 254022, 2011

[56] D. Ielmini, “Resistive switching memories based on metal oxides: mechanisms, reliability and scaling” Semiconductor Science and Technology, Volume 31, Number 6, 2016

[57] M. Bocquet et al., “Robust Compact Model for Bipolar Oxide-Based Resistive Switching Memories”, IEEE Transactions on Electron Devices, vol. 61, no. 3, March 2014

[58] Y. Li et al., "Conductance Quantization in Resistive Random Access Memory", Nanoscale Research Letters 10:420, 2015

[59] A. Ghetti et al., "Comprehensive Analysis of Random Telegraph Noise Instability and Its Scaling in Deca–Nanometer Flash Memories", IEEE Transactions On Electron Devices, Vol. 56, No. 8, August 2009

[60] G. Ghibaudo and T. Boutchacha, "Electrical Noise and RTS Fluctuations in Advanced CMOS Devices" Microelectron. Reliab., 42, pp. 573-582, 2002

[61] M. H. Tsai, T. P. Ma, and T. B. Hook, "Channel-Length Dependence of Random Telegraph Signal in Submicron MOSFETs" IEEE Electron Device Lett., 15, pp. 504-506, 1994

[62] Z. Shi, J-P. Miéville and M. Dutoit, " Random Telegraph Signals in Deep Submicron n-MOSFET’s", IEEE Transactions On Electron Devices, Vol. 41, No. 7, July 1994

[63] J.P. Campbell et al., "Random Telegraph Noise in Highly Scaled nMOSFETs", IEEE 47th Annual International Reliability Physics Symposium, Montreal, 2009

[64] N. Raghavan et al., “Microscopic origin of random telegraph noise fluctuations in aggressively scaled RRAM and its impact on read disturb variability,” in Proc. IEEE International Reliability Physics Symposium (IRPS), April 2013

[65] D. Veksler et al., “Methodology for the statistical evaluation of the effect of random telegraph noise (RTN) on RRAM characteristics,” in Proc. IEEE Int. Electron Devices Meeting (IEDM), December 2012

[66] Z. Fang et al., “Low-frequency noise in oxide-based TiN/HfO_x/Pt resistive random access memory cells,” IEEE Trans. Electron Devices, vol. 59, no. 3, pp. 850–853, March 2012

[67] M.B. Gonzalez et al., "Dedicated random telegraph noise characterization of Ni/HfO_2-based RRAM devices", Microelectronic Engineering 147, pp. 59-62, 2015

[68] Y.T Chung et al., " Investigation of Random Telegraph Noise Amplitudes in Hafnium Oxide Resistive Memory Devices", IEEE International Reliability Physics Symposium (IRPS), 2014

[69] S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy and D. Ielmini, “Statistical Fluctuations in HfOx resistive-switching memory: Part II—Random Telegraph Noise”, IEEE Transactions on Electron Devices, vol. 61, no. 8, august 2014

[70] F.M Puglisi, P. Pavan, A. Padovani, L. Larcher and G. Bersuker, " Random Telegraph Signal noise properties of HfOx RRAM in high resistive state", Proceedings of the European Solid-State Device Research Conference (ESSDERC), 2012

[71] F.M Puglisi, L. Larcher, A. Padovani and P. Pavan, "A Complete Statistical Investigation of RTN in HfO2-Based RRAM in High Resistive State", IEEE Transactions on Electron Devices, Vol. 62, No. 8, August 2015

[72] F.M Puglisi, L. Larcher, A. Padovani and P. Pavan, "Anomalous random telegraph noise and temporary phenomena in resistive random access memory", Solid State Electron, 2016

[73] F.M Puglisi et al., " A Microscopic Physical Description of RTN Current Fluctuations in HfOx RRAM", IEEE International Reliability Physics Symposium (IRPS), 2015

[74] T. Grasser, “Stochastic charge trapping in oxides: from random telegraph noise to bias temperature instabilities”, Microelectronics Reliability, vol. 52 p. 39-70, 2009

[75] F.M Puglisi, P. Pavan and L. Larcher, " Random Telegraph Noise in HfOx Resistive Random Access Memory: from Physics to Compact Modeling", IEEE International Reliability Physics Symposium (IRPS), 2016

[76] D. Ielmini, F. Nardi and C. Cagli, " Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories", Applied Physics Letters 96, 053503, 2010

[77] M. Terai, Y. Sakotsubo, Y. Saito, S. Kotsuji, and H. Hada, " Memory-State Dependence of Random Telegraph Noise of Ta2O5/TiO2 Stack ReRAM", IEEE Electron Device Letters, Vol. 31, No. 11, November 2010

[78] C. Nguyen, C. Cagli, G. Molas, B. Sklenard, C. Nail, K. El Hajjam, J.F. Nodin, C. Charpin, S. Bernasconi and G. Reimbold, “Study of Forming impact on 4Kbit RRAM array performances and reliability”, IEEE 9th International Memory Workshop (IMW), 2017

Conclusion générale et perspectives

Avec l'explosion de l'électronique grand public, les technologies microélectroniques sont en perpétuelle évolution, afin de suivre la cadence imposée par nos besoins, toujours plus avides de performances. Ceci est particulièrement vrai pour les technologies des mémoires, dont le secteur est en pleine effervescence, et plus précisément les technologies de mémoires non-volatiles, particulièrement nécessaires pour des appareils nomades. Depuis des années, et encore aujourd'hui, le secteur des mémoires non-volatiles est dominé par les mémoires flashs. Ces mémoires, dont le fonctionnement repose sur le stockage de charges électriques au sein d'une grille flottante en silicium polycristallin, ont bénéficié des grands progrès en termes de miniaturisation des technologies CMOS, ce qui leur confère des performances remarquables en matière de densité de stockage. Cependant, des contraintes physiques liées à leur intégration dans les nœuds 28nm et en dessous, ont commencé à soulever des questions quant à la succession des mémoires flash. Ainsi, depuis une vingtaine d'année, un certain nombre de technologies de mémoires non-volatiles émergentes a vu le jour, telles que les mémoires magnétorésistives MRAM, les mémoires à changement de phase PCRAM, ou les mémoires résistives RRAM. Ces dernières, sont elles-mêmes divisées en deux catégories : les mémoires dites à pont conducteur CbRAM, et les mémoires résistives à base d'oxydes OxRAM. La différence entre ces deux familles est la nature du filament conducteur à la base de la conduction. Chez les CbRAM, ce filament est créé par la migration d'ions métalliques depuis une électrode active. Chez les OxRAM, il s'avère plutôt que la formation du filament conducteur est liée à la migration d'ions oxygène depuis l'oxyde, laissant un filament composé de lacunes d'oxygène.

Dans ce rapport de thèse, nous nous sommes concentrés sur la famille des OxRAM. Dans la littérature, on peut trouver un grand nombre d'oxydes métalliques pouvant être utilisés comme oxydes pour des OxRAM. On peut ainsi citer l'oxyde d'hafnium, de tantale, de nickel ou bien de titane. Parmi tous ces matériaux, l'oxyde d'hafnium HfO_2 est l'oxyde le plus étudié car il présente de très bonnes performances dans la plupart des secteurs (vitesse, tensions, endurance, miniaturisation ou encore rétention de données), tout en étant complètement compatibles en Back-End-Of-Line des circuits CMOS. C'est la raison pour laquelle nous avons aussi choisi d'étudier des mémoires OxRAM à base d'oxyde d'hafnium.

Si originellement les OxRAM étaient étudiées dans le but de concurrencer puis de remplacer un jour les mémoires flash, cet objectif paraît aujourd'hui plus lointain. En effet, avec l'avènement des 3D-NAND, il semble plus difficile pour les OxRAM d'atteindre les mêmes performances de densité que les mémoires flash. En revanche, les OxRAM se distingue des flash par une vitesse de fonctionnement bien supérieure, capable de concurrencer les mémoires dynamiques DRAM, qui sont pourtant une technologie de mémoires volatiles, réputées bien plus rapide que les mémoires non-volatiles. De plus, leur coût de fabrication est bien inférieur à celui de ces mémoires DRAM. C'est pourquoi, eu lieu de s'intégrer en lieu et place des flashs, on pense aujourd'hui que les OxRAM ont le potentiel pour venir combler le fossé qu'il existe actuellement entre les mémoires flashs (relativement peu rapide, mais avec d'excellentes capacités de stockage) et les mémoires DRAM (très rapide, mais très couteuse et plus difficile à miniaturiser), dans une catégorie qualifiée de « Storage Class Memory ». Cependant, les OxRAM ne peuvent aujourd'hui pas encore être considérée comme une technologie mature. Elles présentent en effet une variabilité trop importante, ce qui conduit à des défaillances en matière de fiabilité. Cet aspect est unanimement considéré comme le principal talon d'Achille des mémoires OxRAM. L'amélioration de cette variabilité passe par une meilleure compréhension de celle-ci, mais également des mécanismes plus généraux de commutation des OxRAM.

Dans cette optique, ce manuscrit de thèse s'attache à comprendre ces mécanismes, via une étude approfondie de la dynamique de commutation des OxRAM à base d'oxyde d'hafnium. L'objet du second chapitre de ce manuscrit est ainsi l'étude expérimentale, via de la caractérisation électrique, de cette dynamique de switching, grâce à un véhicule de test spécial, avec des points mémoires OxRAM avec électrode de titane entièrement fabriqués au sein du CEA-LETI, et à un banc de test entièrement dédié à cette tâche.

Après nous être assuré que notre véhicule de test était fonctionnel, via des mesures « classiques » de fiabilité, à la fois en quasi-statique et en dynamique, nous avons pu commencer les mesures dans le but d'observer « en direct » la commutation résistive dans ces mémoires. Ce type de mesure, très rarement décrite dans la littérature, consiste à placer un point de mesure de tension entre le transistor (qui sert à contrôler le courant qui circule dans le circuit durant les opérations de forming et d'écriture) et la mémoire. Nous avons démontré l'observation de la commutation résistive à la fois sur les opérations d'écriture et d'effacement. Nous avons été capables de détecter ce changement de résistance, particulièrement brusque pour l'opération d'écriture, sur des temps très courts, de l'ordre de la dizaine de nanoseconde.

L'un des grands intérêts de ces mesures, hormis la démonstration de la grande vitesse de commutation des OxRAM, est que nous avons pu déterminer, pour chaque évènement de commutation la tension réelle à la laquelle la mémoire change d'état. Nous avons ainsi pu tracer l'évolution de cette tension de switching en fonction de la vitesse de rampe de tension appliquée (sur l'électrode supérieure pour le set, et l'électrode inférieure pour le reset), et ce, sur une très large plage de vitesse de rampe, ce qui n'avait jamais été réalisé jusque-là sur des structures 1T1R.

Ensuite, nous avons étudié l'impact que la réduction du temps de pulse peut avoir sur les distributions de résistance. Nous avons pu modéliser le comportement de ces distributions, en fonction de ce temps de pulse, et du courant de compliance (pour l'opération d'écriture) ou de la tension de reset. Nous avons ainsi démontré l'intérêt d'utiliser des temps de pulses très courts, puisque le raccourcissement des temps de pulse n'a qu'une influence secondaire sur la dispersion des distributions. En effet, le courant de compliance et la tension de reset utilisés ont une influence bien plus importante. De plus, l'abaissement des temps de pulses, sur plusieurs ordres de grandeurs, a des conséquences très positives d'un point de vue de la consommation d'énergie.

Avec à notre disposition des caractérisations électriques de la dynamique de commutation des mémoires OxRAM, nous avions alors la possibilité de nous interroger sur les mécanismes physiques responsables de ces changements de résistance. Pour cela, le troisième chapitre de ce manuscrit est dédié à la mise au point d'un modèle physique semi-analytique. Ce modèle avait pour objectif de décrire la physique de la commutation, tout en restant suffisamment simple pour être capable de générer des centaines de simulations destinées à être comparées à des mesures de distributions expérimentale. Ce modèle est basé sur le calcul du nombre de lacune d'oxygène générées, ou éliminées (respectivement lors des opérations d'écriture et d'effacement), au sein du filament conducteur qui relie les deux électrodes. Le modèle établie le lien entre les paramètres électriques, comme la résistance du point mémoire ou les tensions et courants de compliance utilisés, avec des paramètres physiques, comme le nombre de lacunes d'oxygène, la taille du gap dans le filament, ou le rayon de celui-ci.

Nous avons ensuite démontré la capacité de notre modèle à fitter des courbes de mesures électriques expérimentales, à la fois en quasi-statique et en dynamique, et pour les opérations d'écritures et d'effacement. Plus particulièrement, nous avons pu reproduire les courbes d'évolution des tensions réelles d'écriture et d'effacement, en fonction de la vitesse de rampe utilisée. Ceci nous a permis d'extraire des paramètres physiques importants, tels que la taille du

gap dans le filament et l'énergie d'activation de génération (ou de destruction) de lacunes d'oxygène. Nous avons montré que ces valeurs dépendaient du matériau choisi comme électrode supérieure, en reproduisant ces courbes pour 4 nouveaux empilements. Nous avons pu comparer les valeurs d'énergie d'activation de génération avec des calculs *ab initio*. Nous avons été par ailleurs capable de déceler une différence de fonctionnement, avec l'empilement avec une électrode de cuivre, qui semble conduire à un mécanisme hybride entre une OxRAM et une CbRAM.

Pour finir, le quatrième chapitre de ce rapport de thèse a pour but de rassembler les connaissances obtenues à l'issue des deux premiers chapitres pour se pencher sur des questions de fiabilité, qui sont, pour rappel, des questions cruciales pour l'avenir des OxRAM et leur possible industrialisation. Pour cela, nous avons réalisé un grand nombre de tests sur des matrices d'OxRAM, également fabriquées par le LETI, afin d'obtenir des informations de statistiques, beaucoup plus élaborées. Nous avons donc commencé ce chapitre par décrire en détail le nouveau banc de test dédié à la caractérisation de matrices mémoires.

Toujours dans une volonté de continuer notre étude liée à la dynamique de commutation, nous avons mesuré les distributions de tensions de switching, pour différentes vitesses de rampes. Ces mesures étaient assez semblables à celles réalisées dans le second chapitre de ce manuscrit, à la différence près qu'il s'agit de tests automatisés sur des milliers de mesures, mais sur des temps plus longs. Les résultats obtenus étaient en accord total avec les mesures précédentes. Ceci nous a permis de les comparer avec des simulations réalisées par notre modèle. En introduisant une distribution normale de la valeur du gap du filament, nous avons pu modéliser ces courbes de distributions.

Par la suite, en mesurant l'impact qu'a le courant de compliance utilisé lors du forming sur ces distributions et en utilisant notre modèle pour reproduire cet impact, nous avons confirmé l'hypothèse souvent utilisée dans la littérature, selon laquelle le courant de forming contrôlait le rayon du filament du conducteur. Ceci nous a amené à réaliser des mesures de bruit RTN sur nos dispositifs. Ce phénomène est un facteur majeur de dégradation de la fiabilité des OxRAM et peut notamment conduire à des erreurs de lectures, mais il permet également de tirer un certain nombre d'informations complémentaires au sujet des mécanismes microscopiques en jeu. Il consiste en des oscillations de courant d'une cellule entre plusieurs niveaux au cours du temps, dues au piégeage de charges par des défauts. La comparaison des mesures de bruits après des opérations de forming d'intensité différente a permis de constater une augmentation du nombre de niveaux de courant à mesure que le courant de forming

augmente, ce qui est cohérent avec une augmentation du rayon du filament. Cependant, si le nombre de niveaux de courant augmente, ce n'est pas le cas de l'amplitude des oscillations (autrement dit, l'écart entre ces niveaux de courants). En effet, augmenter le courant de forming a tendance à faire baisser de façon significative cette amplitude. Il semble ainsi qu'un filament plus épais, résultant d'un courant de forming plus intense est moins sensible à des perturbations provenant de quelques défauts sa proximité. Ceci est très bénéfique d'un point de vue de la fiabilité. En effet, avec des mesures de distributions résistives sur des matrices de 4kbits, nous avons observé que cet effet se ressent dans ces distributions : celles-ci sont ainsi moins dispersées après des opérations de forming plus fortes. En plus de confirmer que le bruit RTN a bien des conséquences sur la variabilité des OxRAM, ceci nous permet de proposer une piste pour améliorer cette dernière.

Pour conclure, à travers cette étude nous avons analysé en profondeur les performances en régime dynamique des mémoires OxRAM, via des techniques de caractérisation de dispositifs unitaires innovantes. Nous avons pu comparer nos résultats expérimentaux à un modèle physique que nous avons mis au point, afin de les relier à des paramètres physiques microscopiques. Ces travaux ont également nécessité l'optimisation et l'utilisation d'un banc de test dédié aux matrices mémoires, afin de croiser les informations très précises et très pointues obtenues sur dispositifs unitaires, avec des informations plus quantitatives sur de la statistique de mesure en dynamique. En effet, de telles mesures sont aujourd'hui nécessaires dans l'étude des mémoires OxRAM, tant le besoin d'en savoir plus sur la variabilité inhérente à cette technologie est important. Ceci nous a au final conduit à analyser des mesures de bruit RTN, c'est-à-dire des mesures d'oscillations au cours du temps du niveau de courant dans les mémoires, qui se sont avérées avoir un réel impact sur les distributions résistives des mémoires étudiées, et nous fournir des informations liées aux évènements microscopiques ayant lieu à proximité du filament.

Il ressort de ces travaux un certain nombre d'informations nouvelles ou complémentaires, qui peuvent chacune mériter d'être approfondies à l'avenir :

- Il est indéniable que les performances en termes de vitesses des mémoires OxRAM sont largement suffisantes pour intégrer le marché. Notre étude nous a conduit à analyser ces performances avec un suivi dynamique sur des temps allant jusqu'à la dizaine de nanosecondes. Si ces temps restent très compétitifs, nous savons que les OxRAM sont capables d'atteindre des temps encore inférieurs. Dans une volonté de mettre au point notre modèle physique, nous

avons choisi de ne pas aller plus loin à ce niveau-là. Néanmoins, il pourrait être intéressant d'explorer les régions de temps inférieures à la dizaine de nanoseconde, pour voir si un changement dans la dynamique de commutation se produit. De plus, nous avons remarqué qu'abaisser les temps de pulses jusqu'à 20ns n'a qu'un impact mineur sur les distributions résistives. Ces travaux pourraient être complétés en regardant si la situation reste la même en passant en dessous des 10ns.

- Nous avons mis en évidence le rôle que joue le bruit RTN dans la dégradation de la fiabilité des mémoires OxRAM, avec notamment l'augmentation de la dispersion des distributions qu'il induit. Il s'est avéré que ces découvertes mériteraient un travail approfondi, uniquement dédié à l'étude de ce bruit RTN. La mise au point d'un traitement des données RTN automatisé semble nécessaire, pour augmenter la statistique de données récoltées, à la fois pour enquêter plus en profondeur sur le rôle du bruit RTN dans la grande variabilité des OxRAM, mais également car ces mesures de bruit sont un outil très efficace pour sonder les évènements à échelle microscopiques que sont les piégeages et dépiégeages de charges par des défauts. Or ces évènements microscopiques pourraient nous en apprendre davantage sur les réels mécanismes physiques des OxRAM, qui ne sont pas encore complètement compris.

Remerciements

Avant tout je tiens à remercier chaleureusement les membres du jury, qui ont accepté d'évaluer ces travaux. Merci donc à Jean-Michel Portal – Aix Marseille Université, et à Damien Deleruyelle – INSA Lyon, pour avoir accepté d'être les rapporteurs de ces travaux. Merci pour leurs commentaires constructifs et intéressés au sujet de mon manuscrit, et pour leurs questions pertinentes lors de la soutenance. Merci également à Anne Kaminski – Grenoble INP, pour avoir accepté de tenir le rôle de présidente du jury. Les commentaires et questions soulevés lors de la soutenance ont également été très constructifs, surtout considérant que les mémoires sont un sujet relativement éloigné de son cœur de métier. Je vous remercie enfin tous les trois, pour l'ambiance chaleureuse que vous avez insufflée à ma soutenance, qui aurait pu, sous d'autres circonstances, être un évènement bien plus stressant pour moi.

Je tiens maintenant à remercier Carlo, mon encadrant CEA. Il était, sans concurrence possible, la personne avec laquelle j'ai eu le plus d'échanges au sujet de mon travail de thèse. Merci pour tous ses conseils avisés et sa grande expertise dans le domaine des mémoires OxRAM. Il est plus qu'évident que ce travail aurait été infiniment moins riche sans sa présence, son assistance et son soutien. Enfin, merci pour sa bonne humeur communicative, son positivisme et son sens de l'humour, qui m'ont permis de travailler dans une ambiance très agréable. Merci à Gérard, mon directeur de thèse. Si son suivi de la thèse était logiquement (de par le fait que je le partageais avec je-ne-sais-combien d'autres thésards) plus distant que celui de Carlo durant ces trois ans, son assistance durant la dernière ligne droite a été cruciale. Par ailleurs, son expérience inégalable et son calme, cumulés à sa grande sympathie, m'ont été très utile pour garder mon sang froid dans les périodes les plus pressantes de la rédaction du manuscrit. Je tiens aussi à remercier Jean, qui, s'il n'a pas son nom sur les documents officiels reliés à ma thèse, a quand même été mon encadrant de stage, durant mon projet de fin d'étude, qui a débouché sur ma thèse. Si j'en suis là, c'est quand même grâce à lui. Je me souviens des discussions avec d'autres thésards ou stagiaires venant d'autres labos, qui m'enviaient d'avoir « le mec le plus sympa possible » comme encadrant de stage. Néanmoins sa grande gentillesse et ses qualités humaines n'étaient pas ses seuls atouts : il m'a également beaucoup appris du point de vue technique, mais c'est moins drôle à signaler dans des remerciements.

Un grand merci à Gilles de m'avoir accueilli au sein du LCTE. Merci de façon plus générale à l'ensemble du laboratoire, dont je n'ai pas le courage de citer tous les membres. Je

pense avoir eu une chance énorme d'avoir réalisé ma thèse au sein d'un groupe si sympathique. Je n'imaginais pas, avant mon stage dans ce groupe, qu'on pouvait trouver une ambiance et une atmosphère si agréable au sein d'un labo de caractérisation et de tests électriques… Merci à vous tous pour ces trois ans en votre compagnie. Petite dédicace à la team Mot-Fléchés du matin. Merci particulier à Estelle. Ça n'a pas dû être toujours facile de gérer mes dossiers, mes missions… Un grand bravo pour sa patience.

Parfaite transition pour remercier mes deux compères Antoine et Thomas (et son alias inexplicable Pascal le Gitan), qui en ont, je crois, également fait voir de toutes les couleurs à Estelle. Merci pour leurs blagues parfois douteuses (surtout Antoine, il faut bien l'avouer), leur enthousiasme, et pour les matches de Soccer 5. Je suis sûr qu'on va continuer à en faire à l'avenir (si je ne me blesse pas durant mes activités de ninja).

J'en profite maintenant pour remercier mes amis grenoblois, qui sont ceux qui partagent la grande majorité de mon temps. David Pagnon, qui n'a pas de cheveux, Fantin, et à la base de données de photos-dossiers laissée sur mon portable, Paulin, qui devrait arrêter de tomber un jour, Noémie, qui n'a probablement pas achevé son incessante transition politique, Arnaud, pour la boxe et la psychologie évolutionniste, Ced, qui va finir par associer le mot backflip à sa passion des cordes, avec la bienveillante aide de Kim, Charlotte, qui est tout simplement la personne la plus appréciée de l'univers, Mathieu, qui n'a toujours pas fini de casser le game des débats FB, Lélia, que je recroiserai sûrement d'ici peu avec grand plaisir, et Déborah, sans qui mes pots de départ n'auraient été que les ombres d'eux-mêmes…

(à noter que là, je mets juste des petites phrases pour chacun, mais vous en mériteriez un manuscrit entier <3)

Merci, à mes grands amis de toujours, Alex, Paul, Ludo, et à ma schreckliche Zozo. Merci à tous les autres dont la liste serait trop fastidieuse à faire ici… (et qui mériterait donc un autre manuscrit…).

Bien entendu, un immense merci à ma famille : papa, maman, Baptiste. Merci pour votre soutien indéfectible, vos petits mots et pour l'accueil inégalable que vous me faites quand je rentre à Angervilliers.

Enfin, merci à ma Sophia. Malgré cette histoire de cheville, et le fait qu'il s'agissait de ma troisième année de thèse, cette dernière année était probablement la plus belle et la plus riche de ma vie, en ta compagnie. Mais je suis sûr qu'on peut faire encore mieux les années suivantes !

Ah ! Et merci, Dimitri, mon ami… <3